An Artificial Intelligence-Based Industry Peer Grouping System

by George Bonne, Andrew W. Lo, Abilash Prabhakaran, Kien Wei Siah, Manish Singh, Xinxin Wang, Peter Zangari, and Howard Zhang

Authors’ final manuscript as accepted for publication

| Citation          | Bonne, George, Andrew W. Lo, Abilash Prabhakaran, Kien Wei Siah, Manish Singh, Xinxin Wang, Peter Zangari, and Howard Zhang (2022), “An Artificial Intelligence-Based Industry Peer Grouping System,” *Journal of Financial Data Science* 4(2), 9–36, |
|-------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| As Published      | https://doi.org/10.3905/jfds.2022.1.090                                                                                                                                                                                                                             |
| Publisher         | Pageant Media Ltd.                                                                                                                                                                                                                                                   |
An Artificial Intelligence-Based Industry Peer Grouping System

George Bonne\textsuperscript{a}, Andrew W. Lo\textsuperscript{b,1}, Abilash Prabhakaran\textsuperscript{c}, Kien Wei Siah\textsuperscript{d}, Manish Singh\textsuperscript{e}, Xinxin Wang\textsuperscript{f}, Peter Zangari\textsuperscript{g}, Howard Zhang\textsuperscript{h}

\textsuperscript{a}George Bonne is Executive Director at MSCI Inc.  
George.Bonne@msci.com
\textsuperscript{b}Andrew W. Lo is the Charles E. and Susan T. Harris Professor at the MIT Sloan School of Management.  
alo-admin@mit.edu
\textsuperscript{c}Abilash Prabhakaran is an U.G. researcher at MIT Laboratory for Financial Engineering  
abilashp@mit.edu
\textsuperscript{d}Kien Wei Siah is a Research Affiliate at MIT Laboratory for Financial Engineering  
kienwei@mit.edu
\textsuperscript{e}Manish Singh is a Ph.D. student at MIT Electrical Engineering and Computer Science  
smanish@mit.edu
\textsuperscript{f}Xinxin Wang is Executive Director at MSCI Inc.  
Xinxin.Wang@msci.com
\textsuperscript{g}Peter Zangari is the Global Head of Research and Product Development at MSCI Inc.  
Peter.Zangari@msci.com
\textsuperscript{h}Howard Zhang is Vice President at MSCI Inc.  
Howard.Zhang@msci.com

\textsuperscript{1}Corresponding author

\textsuperscript{⋆}We gratefully acknowledge the research collaboration between MSCI Inc. and MIT Laboratory for Financial Engineering. We benefited from the methodology of MSCI Peer Similarity Scores product and the data and computing support from the MSCI Data Science Platform. We thank Roman Kouzmenko and Manuel Rueda for insightful discussions. The views and opinions expressed in this article are those of the authors only, and do not necessarily represent the views and opinions of any institution or agency, any of their affiliates or employees, or any of the individuals acknowledged above.
An Artificial Intelligence-Based Industry Peer Grouping System

Abstract
In this work, the authors develop a data-driven peer grouping system using artificial intelligence (AI) tools to capture market perception and, in turn, group companies into clusters at various levels of granularity. In addition, they develop a continuous measure of similarity between companies; use this measure to group companies into clusters and construct hedged portfolios. In the peer groupings, companies grouped in the same clusters had strong homogeneous risk and return profiles, while different clusters of companies had diverse, varying risk exposures. The authors extensively evaluated the clusters and found that companies grouped together by their method had higher out-of-sample return correlation but lower stability and interpretability than companies grouped by a standard industry classification system. The authors also develop an interactive visualization system for identifying AI-based clusters and similar companies.

Highlights
• We developed a data-driven industry peer grouping system that clustered similar companies at different levels of granularity. We also developed an interactive tool to visualize clusters and nearest neighbors of companies.

• Artificial intelligence techniques can extract features from relevant data sources and learn relationships that can identify companies that are similar in terms of risk–return profile for the out-of-sample period.

• Historical returns correlation, GICS classification, 10-K reports, and fundamental factors like size, momentum, and debt-to-asset ratio contributed the most in the predicting the similarity of companies.

Keywords: Machine Learning; Artificial Intelligence; Peer Grouping; Industry Classification; Returns Co-movement

JEL Classification: G10
Industry classification refers to the organization of companies into groups. It has broad application because national agencies need it to summarize economic statistics accurately, and investors and financial analysts use it for investment management and competitive industrial diligence.

Industry classification started in 1937, via the Standard Industrial Classification system (SIC). This system classified companies based on their activities at the establishment level. Since then, multiple industry classification systems have been established, grouping companies by different criteria. For example, the North American Industry Classification System (NAICS) groups companies based on production at the establishment level, the Global Industrial Classification System (GICS\textsuperscript{\textcopyright}2) and the Industry Classification Benchmark (ICB) classify companies according to revenue, earnings and market perception (Methodology, 2020), and the Text-Based Network Industry Classifications (TNIC) groups companies according to business descriptions in 10-K reports.

Currently, none of the existing classification systems are solely based on market perception. However, due to the widespread adoption of industry groupings by multiple stakeholders in investment management, including traders, hedge funds, asset management firms, and stock brokers, the public market’s perception has become vital in classification systems because it allows stakeholders to efficiently compare a company’s financial performance with its peers, and it allows the hedging and construction of diversified portfolios.

In this paper, we created a dynamic industry classification system that groups stocks according to quantified similarities from a wide variety of structured and unstructured data features. We trained machine learning (ML) models of these features to predict future return correlations between stocks, which we view as a proxy for market perception. We define market perception as investors’ views about a company that drive the stock returns’ movement and their co-movement with similar stocks. However, future return co-movement is dynamic and difficult to measure. Nevertheless, with the availability of big data, we were able to use artificial intelligence (AI) methods to extract relevant information about companies from various data sources and learn about their similarity in the future, according to market perception. Our AI-driven Peer Grouping System (AIPGS) adapts itself and evolves over time in tandem with the changing perceptions of companies.

Return correlation between companies has been widely used to determine the homogeneity of industry groups as perceived by investors (Chan et al., 2007; De Bodt et al., 2020). We observed that clusters using a distance metric calculated via return correlation will group companies with similar return characteristics. We calculated the industry effects, and demonstrated that our ML models group firms together with similar fundamental factor exposures. The clusters obtained by this method enhance the potential benefits that can be gained by diversification based on clusters in the portfolio creation process.

In our dynamic peer grouping system, we applied AI methods at different steps to en-
hance the system's ability to group together similar firms. Specifically, we first obtained several datasets on returns, factor exposure, news co-mentions, and 10-K filings. From these datasets, we extracted features via various methodologies, including (but not limited to) term frequency-inverse document frequency (TF-IDF) and Doc2Vec for 10-K filings and Node2Vec for news mentions, to capture the potential similarity between firms\(^3\). We then used ML models such as ridge regression, neural networks, and XGBoost to learn and predict the relationship between features and future return correlation between companies. Finally, we used hierarchical clustering to group similar firms together in sector, industry, and sub-industry groups.

In AIPGS, we learned relationships and found groups that are not captured in existing classification systems. For example, in 2018, one sub-industry group from AIPGS included only Alphabet, Netflix, Apple, Amazon, and Facebook. All of these companies are prominent U.S. technology companies and popularly referred to as FAANG. However, GICS would have traditionally classified these companies in different sub-industries: respectively, as interactive media and services; movies and entertainment; technology, hardware, storage, and peripherals; internet and direct marketing retail; and interactive media and services. Likewise, GICS generally grouped PepsiCo with the Coca-Cola Company and classified PepsiCo as a soft drinks or beverage company at subindustry and industry levels, but AIPGS generally separated PepsiCo from Coca-Cola and grouped PepsiCo in the same subindustry and industry as packaged food or food companies such as General Mills, Mondelez International, and The Hershey Company. This is insightful because PepsiCo earns more revenue from its food/snacks business than from beverages\(^4\).

Similarly, GICS groups Disney in the group movies and entertainment, whereas AIPGS grouped the Walt Disney Company with cable and broadcasting companies such as Comcast Corporation. Disney launched their streaming service Disney+ in 2019, which might potentially explain the grouping. Likewise, Cognizant is an information technology (IT) company that performs data analytics and processing. Although GICS placed Cognizant with IT consulting firms, AIPGS grouped Cognizant with multiple data processing and IT consultancies.

AIPGS also captured groupings that may seem unintuitive or surprising at first glance but can be explained via further fundamental analysis. For example, although eBay may be seen as one of the largest e-commerce sites, AIPGS indicates otherwise. For 2014, the ML model predicted that eBay should have been grouped in the same sub-industry as finance companies such as The Green Dot Corporation, Visa, and Mastercard. Upon further analysis, this classification makes sense because in 2013 eBay owned PayPal, which did not split off until 2015. PayPal’s high growth prospects and leadership in the money-transfer space could have caused the market to view eBay as a fintech firm. Interestingly, for 2018, well after eBay had split from PayPal, AIPGS classified eBay in the same sub-industry as e-commerce sites.

\(^3\)We describe TF-IDF, Doc2Vec, and Node2vec in more details in Section Data and Feature Extraction. These are the AI based feature extraction methods from the text and graph datasets.

\(^4\)We verified this information via 10-K filings of PepsiCo.
such as Priceline.com, Expedia Group, Etsy, and Shutterstock. This indicates the market’s stabilizing perception of eBay as an e-commerce site after its split from PayPal in 2015. Thus, according to our model, for 2014, an investor looking for exposure to companies with similar stock movements as eBay would not have invested in other popular e-commerce sites, but in fintech and payment processing firms such as Visa and Mastercard.

The popular classification systems, such as NAICS, GICS, and ICB, discretely group companies without providing a continuous degree of similarity within and across groupings. In this work, apart from merely grouping companies, our model produces a continuous measure of similarity, which allows us to find the top $k$ peers of a particular company. We demonstrate that single-stock portfolios can be hedged using peers determined via similarity scores.

Our paper contributes to the literature in data-driven classification systems. Hoberg and Phillips (2016) proposed a text-based classification system based on business descriptions from 10-K filings and showed that their method outperformed SIC and NAICS in grouping together firms with similar profitability, sales growth, and market risk. Lee et al. (2015) used internet co-searches from the U.S. Securities and Exchange Commission (SEC) website to find groups of similar firms, assuming that users of the SEC data-retrieval website search for similar firms together to aid their investment process. Although this method incorporated investor perception, it is limited by availability of search data. Dalziel et al. (2018) proposed designing an industry classification system from inter-firm transaction data, but failed to produce any industry groupings due to lack of data. Gay and Karger (2014) and Sprenger and Welpe (2011) proposed identifying peers from patent citation and stock tweets on Twitter, respectively. Both methods were also limited by availability of data: If a company does not write patents in common with patents of other companies or if a company is not mentioned in stock tweets, then the classification system will fail to assign a grouping. Similarly, Zhang and Bonne (2019) proposed a system to quantify company similarity using different data sources (10-K filings, news co-mentions, returns, fundamentals and analyst coverage overlap).

Our analysis builds on the aforementioned methods, and we propose an industry classification model that is explicitly trained to predict the market perception (return correlation) of companies one year into the future. We also incorporated AI methods that led to performance improvements over the baseline. Some of the existing methods suffer from a lack of data availability, and fail to place all public companies into industry groups. Our system is robust because all the datasets used are available (except for news co-occurrence)\textsuperscript{5} for almost all publicly traded U.S. companies.\textsuperscript{6}

One of our goals was to incorporate a baseline classification system that is robust and

\textsuperscript{5}Our industry classification system is still able to assign an industry grouping if the company is missing in the news dataset.

\textsuperscript{6}The companies used in our analysis covered approximately 99% of the free-float-adjusted market capitalization in the U.S.
popular. GICS is a classification system that is “market-oriented” and used by practitioners. ICB is another market-oriented classification system similar to GICS, as documented in Vermorken (2011). Other classification systems are not as relevant to investors; e.g., SIC and NAICS are product-oriented classification systems used by federal statistical agencies to analyze data about the U.S. economy. Fama and French (1997) developed an industry classification system that is based on SIC. Bhojraj et al. (2003) and Chan et al. (2007) compared GICS to NAICS and Fama–French classifications, respectively, and found that GICS more effectively captures stock co-movement, cross-sectional variations in valuation multiples, and other key financial ratios. Hence, we used GICS at various levels of granularity, particularly at the sector, industry, and sub-industry levels, as a benchmark against our classification system. We show that our classification system outperformed GICS: Our system more effectively grouped companies with strong return co-movement and factor exposures.

Next, we show how we calculated a similarity score to group together the most similar companies. These steps are presented in Exhibit 1. First, we calculated the features for all companies from various data sources, as we will describe in Section Data and Feature Extraction. We then calculated the similarity score between each pair of companies. The similarity score between the companies was obtained using ML models, as we will describe in Section Distance Metrics. With these similarity scores, we determined the distance between every pair of companies and applied hierarchical clustering, as explained in Section Clustering, to group all the companies into clusters. Finally, the clusters obtained were evaluated in an out-of-sample period and used for different applications, as we will describe in Section Evaluating Performance.

![Exhibit 1: Flow Chart of Methodology](image)

**Data and Feature Extraction**

We aim to create a classification system that may be useful for investors by incorporating the market’s perception of companies. Consequently, we use multiple data sources to capture the similarity between companies. Inspired by Zhang and Bonne (2019) and MSCI Peer Similarity Scores (MSCI, 2021), we use stock returns, factor exposures, 10-K filings, news co-mentions, and the GICS grouping data in our analysis. We next summarize these datasets and feature extraction methods.

**Returns**

We obtained the daily and monthly returns of all companies. The number of companies and the number of features are described in Exhibit A1 and in Exhibit 2. Our model
incorporated returns because companies belonging to the same industry groups experience similar shocks, which are reflected in stock returns. For example, a regulation change in favor of electric vehicles (EV) is favorable to EV companies, which may cause their stock prices to increase, while potentially hurting automobile firms reliant on gasoline or diesel. Also, we used return correlation to quantify the similarity between companies.

**Factor Exposures**

Multiple factors have been discovered to explain the source of abnormal returns of companies (Harvey et al., 2015). According to the fundamental factor model (Fama and French, 2015), companies with a similar factor exposure tend to have similar return characteristics. The factor exposures used in our models are the descriptor exposures from the MSCI Barra US Total Market Model (Bayraktar et al., 2015) and are listed in the Appendix in Exhibit A1. It is important to note that, in our model, factor exposures that change slowly over time, such as variability in sales, are driven by inputs that change at an annual or quarterly frequency (slow factors), while others, such as short-term reversal, were designed to have higher turnover (fast factors).

**10-K Filings**

10-K filings are annual regulatory reports required by the SEC. We use the business description section of 10-K filings, which includes detailed information about companies’ businesses and operations. Thus, the filings can be used to generate important features to measure the similarity between companies.

These reports are available in text form. To obtain the features used for the similarity model, we use the following methods:

- **TF-IDF**: TF-IDF is used to determine the most important word in documents. Each document is represented by the vector product of term frequency (the proportion of text the term composes in a document) and the inverse document frequency (the number of documents in which the word occurs) for every word in the document. This method of feature extraction has been used by Loughran and McDonald (2011). Once we calculate the TF-IDF features, we obtain a high-dimensional sparse representation of each document. We reduced the number of dimensions to 100 using truncated singular value decomposition (SVD).

- **Doc2Vec**: This approach involves obtaining a document embedding, as explained in Dai et al. (2015). Doc2Vec is an unsupervised learning method, and using the Distributed Bag of Words and Distributed Memory models, we obtained two representations of every 10-K document, Doc2Vec0 and Doc2Vec1. For the Doc2Vec method, we used a publicly available implementation 7 and made appropriate modifications.

---

7https://markroxor.github.io/gensim/static/notebooks/doc2vec-wikipedia.html
**News Co-Mentions**

From Ravenpack, we obtained information about companies mentioned together in news articles. Companies appearing together in news articles can indicate similarity. For example, Facebook and Twitter can be considered similar because they are often mentioned together in articles about social media.

Using the news dataset, we extracted two types of features for companies:

- **Co-occurrence**: For every pair of companies, the Jaccard similarity is calculated, as follows: Let $A$ be the news articles with mention of company $i$ and $B$ be the set of news articles with mention of company $j$. We use $A \cap B / A \cup B$ as their measure of similarity.

- **Node2Vec**: The co-occurrence method fails to capture higher-level connections between companies. For example, if companies $i$ and $j$ are frequently mentioned together in news articles, and if $j$ and $k$ are mentioned together in others, then companies $i$ and $k$ should be similar to some extent even if $i$ and $k$ are never or infrequently mentioned together. Hence, we extract a set of features based on Node2Vec (Grover and Leskovec, 2016). Node2Vec is a graph-embedding model used to determine the transitive relationships between objects. Therefore, we constructed a graph in which companies are nodes and the co-mentioned companies are connected with the edges. The edges were also weighted by the number of times a pair of companies were mentioned together in an article. Using Node2Vec, we obtained a 128-dimensional embedding of the companies.

**GICS**

As mentioned earlier, GICS is an industry-standard hierarchical classification system used by financial firms and investors globally. Within each hierarchy, every public company is classified into a sector, industry group, industry and sub-industry. It is a market-oriented classification system that uses revenue as a key factor in determining a firm’s principal business activity, along with earnings and market perception. Apart from using GICS as a baseline classification system, we also used it as one of the inputs for the similarity prediction model.

**Distance Metrics**

Given these features of companies, we used cosine and ML metrics to calculate the similarity scores between companies. These models processed the data available at the end of year $t$ (the in-sample period) to predict and estimate the distance (based on return correlation) between the companies in year $t + 1$ (the out-of-sample period). This helps as we seek to avoid any look-ahead bias in the similarity score and distance calculation. Details
| Dataset-Features     | Number of Features |
|----------------------|--------------------|
| Returns-Daily        | 252                |
| Returns-Monthly      | 12                 |
| Factors              | 264                |
| 10K-TF-IDF           | 82606              |
| 10K-TF-IDF(SVD)      | 100                |
| 10K-Doc2Vec0         | 200                |
| 10K-Doc2Vec1         | 200                |
| News-Node2Vec        | 128                |
| GICS                 | 4                  |

Exhibit 2: Datasets and Various Features Extracted

about the cosine distance model of similarity prediction are available in the Appendix in the Cosine Distance section.

We used ML models to predict the out-of-sample distance and similarity between the companies. The ML model took the similarity feature vector between a pair of companies as an input. The similarity feature vector was constructed by appending the L1 distance between different feature vectors with the cosine similarity. It is constructed as follows: Given \( f_i^k \) and \( f_j^k \), we constructed \( F_{i,j} = \bigcup_k \{ |f_i^k - f_j^k| , \cos(f_i^k, f_j^k) \} \). Here, \( f \) is a feature vector, and \( i \) and \( j \) are companies. However, for our ML models, \( k \) may represent different datasets: monthly returns, daily returns, 10-K-TF-IDF(SVD), 10K-Doc2Vec, or news Node2Vec features. We also appended the news co-occurrence to \( F_{i,j} \). Finally, the ML model took \( F_{i,j} \) as an input and predicted the future similarity between the companies. We calculated future similarity between the companies using future daily returns correlations.

The different ML models we used for this analysis were ridge regression, neural networks, and XGBoost. Details of the ML models are included in Appendix in the Machine Learning Models section. The training data consisted of \( F_{i,j} \) between all pairs of companies in the sample. For testing year \( t + 1 \) (out of sample), training was performed on year \( t - 1 \), and validation was performed on year \( t \). Once the ML model was trained and hyperparameters were tuned, we used it to obtain the similarity scores between all the pairs of companies for the out-of-sample year \( t + 1 \). After that we performed an evaluation using the predicted similarity score for out-of-sample period \( t + 1 \). We provide further details about the dataset split in the Appendix in Exhibit C1.

Clustering

Once we have the similarity scores between the companies for each year, we can cluster them into groupings of similar companies. These groupings should have 1) homogeneity, in which similar companies should be assigned to the same clusters, and 2) fair balance, in which a majority of the companies should not be assigned to a minority of clusters.
Existing industry classification systems, such as NAICS, GICS and ICB, are hierarchical. The hierarchical structure of the financial markets is also validated by Mantegna (1999) and Dalziel et al. (2018). Hence, we apply hierarchical clustering via an agglomerative (bottom-up) approach. This approach considers all companies as individual clusters. At every subsequent step, the clusters are merged to move up the hierarchy. Hierarchical clustering requires 1) the distance between the pair of observations, and 2) the linkage, which is the method used to merge the clusters upwards in the hierarchy. The similarity score between the pairs of observations is available as described in Section Distance Metrics. We calculated the distance as \((1 - \text{SimilarityScore})^8\). For the linkage, we applied Ward’s method, which minimizes the intra-cluster variance (i.e., the sum of squares from the center) when two clusters are merged. If a large number of companies are assigned to the same cluster, then the intra-cluster variance increases. Hence, using Ward’s method helped the models develop similar-sized clusters.

Using hierarchical clustering, we created groupings that corresponded to sectors, industries, and sub-industries. For a fair comparison with GICS, we approximately matched the number of clusters at each level to GICS groupings, such that if GICS had 11 sectors, then the number of clusters at the sector level for hierarchical clustering was set to 11. However, hierarchical clustering can be used to obtain an arbitrary number of clusters.

Evaluating Performance

We aim to group stocks based on the perception of the market and investors. In the industry classification system, companies belonging to the same group should be similar, and dissimilar otherwise.

Because the market is driven, in part, by investor perception, the similarity between companies can be determined by calculating the co-movement of stocks. If market participants consider a set of companies to be closely related, those companies may experience similar movements in their stock returns, given similar exposure to risk factors and shocks that influence their returns. The use of return correlation to measure the similarity between companies has been explored earlier by De Bodt et al. (2020) and Chan et al. (2007). We define the return correlation between the companies as a measurement of similarity.

For two stocks, \(i\) and \(j\), at the end of year \(t\), the similarity is given by \(\rho_{i,j}\) where \(\rho\) is the correlation between the daily returns of stocks in year \(t + 1\). To evaluate the goodness of the prediction of \(k\) peers of company \(i\), we calculated the average future returns correlation(\(\bar{\rho}_i\))

\[^8\text{sqrt}(2 * (1 - \text{SimilarityScore}))\) is used in literature to convert correlation to distance. We observed empirically that our transformation method led to similarly sized clusters, with higher values of the average return correlation between companies belonging to the same cluster.
given by: $\bar{\rho}_i = \frac{\sum_{i \neq j} \rho_{i,j}}{k - 1}$

After obtaining the average future returns correlation for each company, we obtain the goodness of the prediction as follows:

$$\bar{\rho} = \frac{\sum_{i=1,K} \bar{\rho}_i}{K}$$

where $K$ represents the total number of stocks being evaluated, and $\bar{\rho}$ represents the overall average return correlation between similar stocks.

A higher value of $\bar{\rho}$ reflects better performance by the classification system in grouping similar stocks. In our evaluation, we used the $\bar{\rho}$ to assess the industry grouping models. While evaluating $\bar{\rho}$, we set values of $k = 2, 5, 10$, and also to ‘dynamic’. Setting $k$ as dynamic implies that the number of peers is decided by the number of companies in the sector, industry or sub-industry cluster to which company $i$ belongs.

We evaluated the ability of the ML models to predict the peers of every company in our dataset. To do this, we evaluated the forward return correlation with the top $k$ peers as predicted by the model. To compare against GICS, we decided the value of $k = dynamic$ for a company by the number of peers in the GICS grouping to which it belongs. For example, at the industry level, if company $i$ belongs to a GICS group that contains $m$ peers, then $k = dynamic$ at the industry level will include $m$ peers in the calculation of $\bar{\rho}_i$. We evaluated peers at three levels: sector, industry and sub-industry. We evaluated the models for the years 2014-2020, and report the average values over all years in the study period.

Exhibit 3 shows the effectiveness of the ML models in predicting the top $k$ peers of each company. The ridge regression model was evaluated after being trained with features from individual datasets. We observed that the ridge regression model with all features (Ridge:ALL) outperformed the ridge regression model trained on only a subset of features. Adding GICS features to the model (Ridge:ALL+GICS) led to small improvements, which implies GICS features have information not captured by other data sources. We see will see later that adding GICS features also increased cluster stability. We also used nonlinear ML models, such as neural networks and XGBoost. In comparison to the ridge regression model, the nonlinear models showed greater peer-prediction power. We also include the performance of different ML models over each year individually in the Appendix in Exhibit D3, and we observed that ML models consistently outperformed GICS at the sector, industry, and sub-industry levels. In Exhibit D5 in the Appendix, we also include the standard errors for the forward returns correlation, using bootstrapping training data for the ridge regression model. We observed that standard errors were smaller than forward return correlation improvements by orders of magnitude. Hence, these improvements were statistically significant.

---

9The first peer of a firm $i$ is firm $i$ itself, hence we don’t include that in the numerator and denominator.
| Model:Features | Sector | Industry | Sub-industry | Top 10 | Top 5 | Top 2 |
|---------------|--------|----------|--------------|--------|-------|-------|
| GICS          | 33.17  | 41.97    | 44.57        | -      | -     | -     |
| Ridge:Factors | 33.69  | 39.99    | 42.13        | 40.80  | 42.16 | 44.15 |
| Ridge:10k-ALL | 33.62  | 40.66    | 43.50        | 41.54  | 43.38 | 46.02 |
| Ridge:Returns | 35.57  | 42.40    | 44.56        | 44.65  | 46.47 | 49.12 |
| Ridge:ALL     | 36.20  | 43.75    | 46.22        | 46.16  | 48.35 | 51.55 |
| Ridge:ALL+GICS| 36.21  | 43.96    | 46.48        | 46.56  | 48.79 | 51.96 |
| NN:ALL        | 36.13  | 43.64    | 46.09        | 46.30  | 48.48 | 51.65 |
| NN:ALL+GICS   | 36.30  | 43.87    | 46.31        | 46.63  | 48.80 | 51.96 |
| XGBoost:ALL   | 36.48  | 44.01    | 46.48        | 46.60  | 48.66 | 51.59 |
| XGBoost:ALL+GICS | **36.58** | **44.26** | **46.69** | **46.84** | **48.95** | **51.96** |

Exhibit 3: Evaluation of ML models. This table presents the values of average return correlation of peers in the out-of-sample periods. The peers are obtained via the ML models for different datasets. For a fair comparison with GICS, the number of peers for sector, industry, and sub-industry groupings was decided according to the number of companies in the GICS group to which the company belonged. ML models trained on all features outperformed GICS. However, the marginal performance improvements gained by adding GICS features (Ridge:ALL+GICS) and by the use of nonlinear ML models (NN:ALL, XGBoost:ALL) were small compared to ridge regression (Ridge:ALL). NN = neural network.

We have included the performance of peers obtained from the cosine distance metric in Exhibit D1 in the Appendix. On comparing cosine and ML models, we observed a minor improvement over GICS classification when using the cosine distance of features to find the top $k$ peers. However, using the ML model’s prediction as a similarity metric improved the return correlation between predicted peers by 2-3 percentage points over the GICS baseline and the best cosine model. This signifies that the ML models more effectively prioritized features that captured the market perception of companies than the cosine distance, which weights all features with equal importance. Thus, having unequal feature importance learned in a data-driven way led to improvements in performance.

After evaluating the peer-prediction power of the model, we next evaluated the clusters obtained after hierarchical clustering. From the predicted correlations, we computed the pairwise distance between companies and used that for hierarchical clustering, as described in the Clustering section. Exhibit 4 presents the average return correlations for clusters at all three levels: sector, industry, and sub-industry. We determined the number of peers in the process of calculating average return correlation (the top $k$) based on the number of companies in the cluster.\(^\text{10}\) We observed that ML clusters outperformed GICS by a large margin (up to 6 percentage points at the sub-industry level and 4 percentage points at the industry level). Adding GICS features as an input to the ML model also increased performance. The best-performing ML clusters were obtained from the ridge and XGBoost models trained with all features, including GICS. We include the performance distribution

\(^{10}\)Note that in Exhibit 3, $k$ was decided based on number of companies belonging to the same GICS group irrespective of model, while in Exhibit 4, $k$ was decided based on the number of companies belonging to the cluster from the corresponding model.
over all years after clustering for different models in Exhibit D4 in the Appendix. The improvements using ML models were consistent over the years. The clusters obtained from the different ML models were the different versions of AIPGS. In subsequent analysis, we observed that different versions had different desirable properties.

| Model:Features | Sector | Industry | Sub-industry |
|----------------|--------|----------|--------------|
| GICS           | 33.17  | 41.97    | 44.57        |
| Ridge:ALL      | 32.04  | 44.10    | 48.75        |
| Ridge:ALL+GICS | **33.54** | **44.73** | **50.07**    |
| NN:ALL         | 30.70  | 41.09    | 48.20        |
| NN:ALL+GICS    | 30.74  | 42.56    | 48.53        |
| XGBoost:ALL    | 32.24  | 43.07    | 49.45        |
| XGBoost:ALL+GICS| 32.90 | 44.69    | **50.64**    |

Exhibit 4: Results after hierarchical clustering. This table presents the average return correlation for clusters obtained by hierarchical clustering on the distance obtained from the ML models. Ridge:ALL refers to the ridge regression model with all the features from various datasets. +GICS refers to the addition of GICS features as one of the inputs to the model. The clusters at the industry and sub-industry levels had higher average return correlations between companies by up to 2.8 and 6 percentage points, respectively, compared to GICS clusters.

The clusters obtained from the ML methods grouped companies with similar market perception and higher return correlations. Apart from obtaining these groups, the ML models also provided a similarity score that can be used to obtain the top $k$ peers for any company. Next, we evaluated the different properties of clusters that are relevant to an industry classification system and measured their potential for diversification and hedging applications.

**Similarity to GICS**

Apart from the source of revenue, GICS also recognizes market perception as relevant for its classification purposes. Our aim is to create company groupings that incorporate market perception more effectively than GICS. Hence, we compared the similarity of our ML clusters to GICS groupings. We calculated this similarity using the adjusted Rand index (ARI), counting pairs of companies assigned to the same and different clusters in GICS groupings and ML clusters, normalized by the total pairs of companies.\(^{11}\) Thus, the ARI represents the probability that a pair of companies will be grouped together by both GICS and our ML clusters. Higher ARI values imply greater similarity between the GICS groupings and ML clusters. Exhibit 5 presents the ARI for different ML groupings at various levels. We observed that without using GICS input features, ML clusters were approximately 30%-50% similar to GICS, and saw the greatest similarity at the industry level. Because GICS is a market-oriented classification system, some degree of similarity between the ML clusters and

\(^{11}\)https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html
GICS is expected. Adding GICS features as an input increased the similarity to the GICS groupings; specifically, the similarity reached up to 71% for the ridge regression clusters.

| Model:Features  | Sector-ARI | Industry-ARI | Sub-industry-ARI |
|-----------------|------------|--------------|------------------|
| Ridge:ALL       | 32.81      | 48.45        | 44.91            |
| Ridge:ALL+GICS  | 58.59      | 71.10        | 63.12            |
| NN:ALL          | 26.15      | 39.18        | 39.09            |
| NN:ALL+GICS     | 28.17      | 42.85        | 41.40            |
| XGBoost:ALL     | 33.02      | 44.73        | 43.62            |
| XGBoost:ALL+GICS| 51.43      | 59.19        | 50.55            |

Exhibit 5: Adjusted Rand Index for ML Clusters. This table presents the ARI values for ML and GICS clusters. Ridge:ALL refers to the ridge regression model with all the features. +GICS refers to addition of GICS features as one of the inputs. We observed that ML clusters had similarities to GICS clusters—an expected result, given that GICS is a market-oriented classification model.

| Granularity      | GICS | Ridge | Ridge+GICS | NN  | NN+GICS | XGB  | XGB+GICS |
|------------------|------|-------|------------|-----|---------|------|----------|
| Sector: 90%      | 97   | 35    | 24         | 41  | 31      | 30   | 26       |
| Industry: 90%    | 97   | 54    | 40         | 24  | 29      | 20   | 32       |
| Sub-industry: 90%| 96   | 49    | 31         | 24  | 27      | 21   | 23       |
| Sector: 50%      | 100  | 95    | 89         | 81  | 91      | 80   | 90       |
| Industry: 50%    | 100  | 69    | 88         | 63  | 69      | 63   | 76       |
| Sub-industry: 50%| 99   | 62    | 75         | 56  | 60      | 54   | 65       |

Exhibit 6: Cluster Stability. In this table, we present the percentage of clusters retaining 90% and 50% of the companies over the years at different granularities of grouping. Ridge refers to the model trained on the all the features. Ridge+GICS refers to the ridge regression model trained on all features, including GICS features. We observed that up to 50% of clusters retained 90% of companies for ridge regression clusters, whereas up to 90% clusters retained 50% of companies for the majority of the models. Adding GICS features in different models generally increased the stability of the clusters. We concluded that, despite the dynamic nature of the clusters, the clusters were fairly stable, because the majority of them mapped from year $t$ to year $t+1$. GICS had the maximum stability. XGB = XGBoost.

**Cluster Stability**

We aim to create a dynamic classification system where the clusters update every year in tandem with the evolution of a company’s businesses and market perception. However, cluster composition should not be drastically different from year $t$ to year $t+1$; major changes in all the companies in our universe are unlikely, and stability is an important and desirable characteristic of a classification system. Hence, we analyzed the stability of clusters across different years. To compare their stability, we calculated the percentage of clusters that retain more than 90% and 50% of companies from year $t$ to year $t+1$. We present the results in Exhibit 6, in which we observed that up to 50% of clusters retained 90% of companies for ridge regression clusters, while up to 90% of clusters retained 50% of companies for a majority of the models. This showed that the ML clusters were not only dynamic, but
fairly stable, and the majority of them were mapped from year \( t \) to year \( t+1 \). Adding GICS as one of the input features increased the stability of the clusters. The clusters from the ridge regression model were more stable than those from XGBoost or neural networks. We also observed that GICS grouping had the maximum stability which is a desirable property of a classification system.

**Cluster Factor and Industry Effects**

We expect that companies belonging to the same clusters will react similarly to different systematic factors. Hence, each cluster of companies can be considered as a factor that can explain the systematic shocks common to the group. We analyzed the clusters obtained from the ML model as factors and evaluated their homogeneity and out-of-sample diversification benefits. We followed the methodology of Heston and Rouwenhorst (1995) to calculate the industry effects, which involves calculating the cluster factor returns using the following cross-sectional regression for every trading day \( t \):

\[
R_{j,t} = A(t) + \sum_{i=1}^{N} L_{i,t} D_{j,i,t} + \epsilon_{j,t}
\]

where \( R_{j,t} \) is the return of stock \( j \) at time \( t \), \( D_{j,i,t} \) is an indicator variable which is equal to 1 if stock \( j \) belongs to cluster \( i \) in time \( t \), and \( \epsilon_{j,t} \) is the return specific to stock \( j \) at time \( t \). \( L_{i,t} \) is the return of cluster \( i \) at time \( t \), which is learned by the model. \( A(t) \) is the intercept for the aforementioned cross-sectional regression. We performed a weighted regression using the square root of market capitalization to give more, but not excessive, weight to companies with high market capitalization, following the methodology of the MSCI Barra US Total Market Model.

| Granularity       | GICS | Ridge | Ridge+GICS | NN | NN+GICS | XGB  | XGB+GICS |
|-------------------|------|-------|------------|----|---------|------|----------|
| Sector            | 0.73 | 0.85  | 0.80       | 0.85| 0.85    | 0.87 | 0.84     |
| Industry          | 0.93 | 1.15  | 1.13       | 1.12| 1.14    | 1.13 | 1.09     |
| Sub-industry      | 1.21 | 1.54  | 1.56       | 1.39| 1.43    | 1.36 | 1.38     |

Exhibit 7: Diversification Table for Cluster Factor Returns. This table presents the standard deviation of cluster factor returns (Equation 1) averaged over time for clusters at different levels of granularity. The cross-sectional dispersion of ML clusters was up to 30% higher than GICS clusters. A higher standard deviation for ML cluster returns implies greater diversification potential using cluster factors and greater power in explaining the cross section of stock returns.

Homogeneity of clusters in terms of return characteristics is verified using forward return correlation. Here, we seek to evaluate the diversification benefits of ML cluster factors. We calculated the cross-sectional dispersion (the standard deviation) for cluster factor returns, as suggested by Solnik and Roulet (2000). A higher cross-sectional dispersion implies greater power in explaining the cross section of security returns, which may help investors as they seek to build better and more efficient portfolios when used in the context of a factor model.
and optimizer.

\[ D_t = \text{std}_i(L_{i,t}) \]  

We averaged \( D_t \) and present the results in Exhibit 7. The cross-sectional dispersion of ML clusters was up to 30% higher than that of GICS clusters. Hence, the ML clusters showed a potential benefit in portfolio construction over GICS.\(^\text{12}\)

To analyze the ability of ML models to group companies with similar fundamentals, we also evaluated the fundamental characteristics of companies grouped together as measured by their size, leverage, profitability and so on. We calculated the fundamental factor exposures for different clusters by averaging the factor exposures of companies belonging to the cluster and calculated the diversification benefits for each fundamental factor using the cross-sectional standard deviation of a cluster’s factor exposure, similar to Equation 1. Along with the inter-cluster dispersion of fundamental factor exposures, we calculated the intra-cluster dispersion of exposures by calculating the standard deviation of factor exposures of companies in a cluster and averaging the standard deviations of all clusters. Smaller values of intra-cluster dispersion signify that companies belonging to the same cluster have similar fundamental values.

Exhibits 8 and 9 present the intra- and inter-cluster dispersion for market beta, dividend yield, earnings yield, growth, size, and residual volatility for clusters at different levels. For factors such as residual volatility, dividend yield, and size, the intra-cluster dispersion was smaller than it was for GICS clusters. Similarly, the inter-cluster variances of factor exposures for residual volatility, size, and earnings yield were 40% higher than they were for the GICS clusters. Lower values of intra-cluster dispersion for ML clusters imply that the ML models grouped together companies with similar fundamentals, and higher values of inter-cluster dispersion imply that ML clusters have greater diversification potential for portfolios. Because fundamental factors are one of the inputs to the ML models, the models learned to group together companies with similar fundamentals. Hence, along with grouping companies with similar return characteristics, the ML clusters also grouped companies with similar fundamental factor exposures.

**Hedging Risk Exposure**

Prior to clustering, we obtain a list of similar stocks, also known as peers, for each company, using both the cosine and the ML models. These peers have similar risk exposures and return characteristics because they have higher values of future returns correlations. Hence, the peers can be used for hedging a portfolio’s risk exposure. We demonstrate this by constructing single-stock portfolios. For every company \( i \) in our universe, we constructed a U.S.-dollar-neutral portfolio by taking a long position on company \( i \) and an equally weighted short position on its top \( k \) peers. We determined the top \( k \) peers using different peer-prediction models, which were either the cosine or ML model. For this analysis, we used ML

\(^{12}\)Although portfolios built using ML clusters may incur higher turnover costs owing to their lower stability compared to GICS.
| Granularity   | GICS | Ridge | Ridge+GICS | NN | NN+GICS | XGB | XGB+GICS |
|---------------|------|-------|------------|----|---------|-----|----------|
| **Market Beta** |      |       |            |    |         |     |          |
| Sector        | 0.82 | 0.79  | 0.79       | 0.76 | **0.76** | 0.76 | 0.78     |
| Industry      | 0.70 | 0.72  | 0.72       | **0.69** | 0.70 | 0.70 | 0.70     |
| Sub-industry  | **0.65** | 0.67  | 0.66       | 0.66 | 0.66 | 0.66 | 0.66     |
| **Dividend Yield** |      |       |            |    |         |     |          |
| Sector        | 0.84 | 0.75  | 0.76       | 0.75 | 0.74 | **0.72** | 0.75 |
| Industry      | 0.71 | 0.68  | 0.68       | **0.66** | 0.66 | 0.67 | 0.68 |
| Sub-industry  | 0.70 | 0.62  | 0.62       | 0.60 | 0.60 | 0.60 | **0.59** |
| **Earnings Yield** |      |       |            |    |         |     |          |
| Sector        | 0.81 | 0.79  | 0.79       | 0.78 | 0.76 | **0.76** | 0.77 |
| Industry      | **0.68** | 0.74  | 0.72       | 0.68 | 0.69 | 0.69 | 0.70 |
| Sub-industry  | **0.63** | 0.68  | 0.68       | 0.66 | 0.67 | 0.67 | 0.67 |
| **Growth**    |      |       |            |    |         |     |          |
| Sector        | 0.91 | 0.93  | 0.93       | 0.91 | **0.90** | 0.92 | 0.93     |
| Industry      | **0.79** | 0.88  | 0.86       | 0.85 | 0.86 | 0.86 | 0.87     |
| Sub-industry  | **0.75** | 0.81  | 0.82       | 0.82 | 0.83 | 0.83 | 0.83     |
| **Residual Volatility** |      |       |            |    |         |     |          |
| Sector        | 0.79 | 0.76  | 0.77       | 0.71 | 0.72 | **0.71** | 0.74 |
| Industry      | 0.71 | 0.73  | 0.72       | **0.66** | 0.67 | 0.67 | 0.68 |
| Sub-industry  | 0.66 | 0.68  | 0.67       | 0.64 | 0.65 | **0.64** | 0.65 |
| **Size**      |      |       |            |    |         |     |          |
| Sector        | 0.98 | 0.94  | 0.96       | **0.90** | 0.91 | 0.90 | 0.92     |
| Industry      | 0.93 | 0.81  | 0.87       | 0.78 | 0.79 | **0.78** | 0.81 |
| Sub-industry  | 0.88 | 0.71  | 0.75       | 0.68 | 0.71 | **0.66** | 0.68 |

Exhibit 8: Intra-Cluster Dispersion for Various Fundamental Factors. The above table presents the intra-cluster dispersion of factor exposures for various clusters. Different columns refer to different ML methods. For example, Ridge refers to the clusters obtained using ridge regression on all features, excluding GICS. Ridge+GICS refers to the ridge regression model trained on all features, including GICS features. For factors such as size, residual volatility and dividend yield, the intra-cluster dispersion for ML clusters was smaller than GICS clusters at all levels. Lower values of intra-cluster dispersion imply that companies in clusters have similar fundamental characteristics.
| Granularity   | GICS | Ridge | Ridge+GICS | NN  | NN+GICS | XGB | XGB+GICS |
|--------------|------|-------|------------|-----|---------|-----|---------|
| **Market Beta** |      |       |            |     |         |     |         |
| Sector       | 0.69 | 0.77  | 0.74       | 0.81| 0.77    | 0.80| 0.76    |
| Industry     | 0.68 | 0.69  | 0.70       | 0.73| 0.74    | 0.73| 0.72    |
| Sub-industry | 0.72 | 0.74  | 0.75       | 0.76| 0.75    | 0.75| 0.75    |
| **Dividend Yield** |      |       |            |     |         |     |         |
| Sector       | 0.60 | 0.73  | 0.68       | 0.76| 0.75    | 0.76| 0.73    |
| Industry     | 0.66 | 0.66  | 0.65       | 0.73| 0.72    | 0.73| 0.70    |
| Sub-industry | 0.73 | 0.69  | 0.68       | 0.71| 0.71    | 0.72| 0.71    |
| **Earnings Yield** |      |       |            |     |         |     |         |
| Sector       | 0.44 | 0.66  | 0.62       | 0.69| 0.66    | 0.71| 0.67    |
| Industry     | 0.50 | 0.67  | 0.67       | 0.70| 0.69    | 0.69| 0.67    |
| Sub-industry | 0.57 | 0.78  | 0.78       | 0.78| 0.78    | 0.79| 0.78    |
| **Growth**   |      |       |            |     |         |     |         |
| Sector       | 0.34 | 0.41  | 0.38       | 0.40| 0.38    | 0.40| 0.39    |
| Industry     | 0.45 | 0.48  | 0.46       | 0.50| 0.48    | 0.47| 0.46    |
| Sub-industry | 0.53 | 0.63  | 0.63       | 0.61| 0.60    | 0.57| 0.57    |
| **Residual Volatility** |      |       |            |     |         |     |         |
| Sector       | 0.49 | 0.66  | 0.63       | 0.72| 0.72    | 0.73| 0.69    |
| Industry     | 0.53 | 0.68  | 0.68       | 0.70| 0.71    | 0.70| 0.69    |
| Sub-industry | 0.61 | 0.78  | 0.79       | 0.78| 0.79    | 0.78| 0.79    |
| **Size**     |      |       |            |     |         |     |         |
| Sector       | 0.22 | 0.32  | 0.23       | 0.37| 0.34    | 0.36| 0.30    |
| Industry     | 0.44 | 0.58  | 0.49       | 0.62| 0.58    | 0.63| 0.57    |
| Sub-industry | 0.62 | 0.70  | 0.68       | 0.73| 0.69    | 0.73| 0.72    |

Exhibit 9: Diversification Table. This table presents the standard deviation of a cluster’s fundamental value averaged over time for clusters at different granularity. Different columns refer to different ML methods. For example, Ridge refers to the ridge regression model trained on all features, excluding GICS. Ridge+GICS refers to ridge regression trained on all features, including GICS. For factors such as residual volatility, size, and earnings yield, the cross-sectional standard deviation was around 40% higher than GICS clusters. Higher standard deviation of fundamental factor exposures for ML clusters implies higher cross-sectional dispersion in cluster factors.
models that were trained on features calculated from all datasets combined. We performed a comparison of different models with their GICS peers. From the GICS grouping, we obtained a set of companies belonging to the same GICS industry. From the same industry cluster, we selected the top \( k \) peers on the basis of size, as done in the literature (Albuquerque, 2009). The above methodology is inspired by Zhang and Bonne. (2020)

After constructing these U.S.-dollar-neutral portfolios, we calculated their daily returns and the value of their annualized volatility. We then averaged the realized volatility for all single-stock portfolios, as shown in Exhibit 10. If the peers have similar risk exposures and return co-movements, we expect the realized volatility of the hedged portfolios to be smaller. Based on Exhibit 10, using one and five peers for hedging single-stock portfolios decreased the realized volatility by 643 and 259 basis points, respectively, when compared to size-based peers obtained from GICS groups.

| Model | GICS | Ridge | Ridge+GICS | NN | NN+GICS | XGB | XGB+GICS | Change |
|-------|------|-------|------------|----|---------|-----|----------|--------|
| \( k = 1 \) | 43.88 | 38.88 | 38.23 | 37.81 | 37.67 | 37.49 | \textbf{37.45} | 643 |
| \( k = 5 \) | 34.21 | 32.02 | 31.91 | 31.90 | 31.81 | 31.70 | \textbf{31.62} | 259 |

Exhibit 10: Hedging Risk Exposures. We present the average annual realized volatility (as percentages) of hedged single-stock portfolios when peers were decided based on different models. The rows correspond to the number of companies selected for hedging. The Change column refers to the difference in the realized volatility of GICS and the best model (XGB+GICS). The volatility was reduced by up to 643 and 259 basis points for \( k = 1 \) and \( k = 5 \), respectively, when peers were obtained from the XGBoost model.

**Model Interpretation**

We have evaluated the abilities of different ML models against GICS groupings to predict the similarity of companies and cluster them effectively. Here, we interpret the predictions of these ML models and find the top-contributing features and datasets that captured the similarity between companies.

We observed that XGBoost was the best-performing model for various metrics. It outperformed ridge regression by a small margin in the average return correlation. For ridge regression, calculating the contribution of features was trivial, given the linearity of the model. For XGBoost, we used the post-hoc method of interpretability, TreeSHAP (Lundberg et al., 2018), which calculates the Shapley values of the features that contribute to the prediction of the output.

We present some examples of predicted peers and the contribution of their features. Exhibit 11 presents the top-10 peers for eBay predicted for the year 2014 by the XGBoost model. We include the predicted correlation and the actual correlation between eBay and its peers. The first six columns show the contributions of the different dataset features. We have highlighted two peers, The American Express Company (AmEx) and IBM. Even though the
GICS contribution for AmEx was negative, the XGBoost model determined AmEx to be a peer due to its high similarity in factor exposures and 10-K filings. We observed that the actual correlation between AmEx and eBay was high in 2014. In the case of IBM, the actual correlation was not high. However, there appears to have been significant contribution by news co-occurrence features. This implies that IBM was mentioned with eBay multiple times in news articles, which led to it being predicted to be one of eBay’s peers. On further analysis, we found that the news articles mentioning eBay and IBM together were earnings announcements on the same dates and did not convey much information about the similarity between the companies. Similarly, we can analyze the feature contributions for individual predictions of all companies by using TreeSHAP for both XGBoost and the model coefficients for ridge regression.

Exhibit 11: Predictions of eBay Peers for 2014. This plot presents the contribution of features from different datasets for the prediction of the top-10 peers of eBay for the year 2014. Features with a high contribution are colored with a stronger degree of red, while features with a weaker contribution are colored by blue. The high similarity between AmEx and eBay can be attributed to the factor exposure and 10-K-TFIDF datasets, even though they belonged to different GICS groups. For the predicted correlation (PRED CORR) and observed correlation (CORR), we colored AmEx green, because it is shown to be the most similar to eBay, and colored IBM red, because it had the least actual similarity, despite being seen as one of eBay’s top-10 peers. D2V=Doc2Vec.

| RET | GICS  | Factor | 10K-TFIDF | 10K-D2V | NEWS | Top-10 Peers | PRED CORR | CORR |
|-----|-------|--------|-----------|---------|------|-------------|-----------|------|
| 0.016 | -0.004 | 0.073 | 0.007 | 0.000 | 0.005 | DANAHER CORP DEL | 0.436 | 0.288 |
| -0.021 | 0.041 | 0.039 | 0.009 | 0.014 | 0.014 | YAHOO INC | 0.437 | 0.383 |
| 0.001 | 0.026 | 0.050 | 0.017 | 0.003 | 0.000 | TOTAL SYS SVCS INC | 0.437 | 0.374 |
| -0.009 | 0.019 | 0.054 | 0.000 | -0.006 | 0.001 | INTERNATIONAL BUSINESS SECURITIES | 0.438 | 0.191 |
| -0.025 | 0.052 | 0.067 | 0.014 | -0.006 | 0.000 | J2 GLOBAL COMMUNICATIONS | 0.442 | 0.338 |
| -0.008 | -0.004 | 0.058 | 0.055 | -0.002 | 0.016 | AMERICAN EXPRESS CO | 0.454 | 0.425 |
| -0.003 | 0.020 | 0.057 | 0.038 | -0.007 | 0.020 | VISA INC | 0.466 | 0.341 |
| 0.000 | 0.033 | 0.043 | 0.042 | 0.011 | 0.002 | VANTIV INC | 0.471 | 0.236 |
| 0.019 | 0.030 | 0.081 | 0.013 | -0.004 | 0.001 | FIDELITY NATL INFORM | 0.480 | 0.383 |
| 0.011 | 0.030 | 0.041 | 0.046 | 0.001 | 0.022 | MASTERCARD INC | 0.492 | 0.346 |

Next, we calculate the global-feature importance over the years for the ridge regression model. We aggregated the feature importance (i.e., the absolute value of a coefficient) over different years, and present the top 10 contributing features in Exhibit 12. We observed that the past returns correlation was the most significant feature, followed by the GICS sub-industry and industry features. The ridge regression model gave maximum importance to the return correlation from historical data, suggesting that the market perception (return correlation) of companies did not drastically change over the years. GICS features were also some of the top contributing features, which was expected, because GICS is a market-oriented classification system, and we are aiming to use market perception to group companies. Feature importance, as presented in Exhibit 12, shows the variables that had the largest impact on the predicted forward return correlation.

We also grouped individual features by their datasets, and present the overall feature importance in Exhibit 13. We observe that the returns dataset had the highest contribution, followed by factor exposure and 10-K information.
Exhibit 12: Feature Importance. The plot shows the 10 most significant features for the ridge regression model aggregated over testing years. The return correlation as calculated from historical data was most important, followed by certain GICS features. Some fundamental factors such as size, momentum, and the debt-to-asset ratio were also found in the 10 most important features.

Exhibit 13: Feature Importance Aggregated by Dataset. The plot shows the importance of features obtained by the absolute value of coefficients from different datasets and aggregated over years for the ridge regression model. The returns dataset was most important, followed by factor exposure and 10-K information.

Cluster Visualization

In our sample, we grouped approximately 2,000 companies into a different number of clusters every year, depending on their sector, industry, or sub-industry level. This makes analyzing and visualizing peers of these companies and classification a non-trivial task. Hence,
we developed a cluster visualization system with multiple functionalities; this allowed us to visualize the companies and their peers in the cluster groups and their nearest neighbors.

Given the distance matrix $D$, $D_{i,j}$ is the distance between companies $i$ and $j$ as predicted by the ML models. We aim to learn a 2D or 3D vector $X_i$ and $X_j$ for companies $i$ and $j$, which approximates $distance(X_i, X_j)$ as $D_{i,j}$. To do this, we used $t$-distributed stochastic neighbor embedding ($t$-SNE) (Van der Maaten and Hinton, 2008). This constructs a probability distribution over pairs of companies in high- and low-dimensional space. It learns the lower dimensional coordinates by minimizing the KL divergence between the probability distributions in low- and high-dimensional space. To visualize the clusters, the lower-dimensional vector $X$, obtained from $t$-SNE, can be plotted for the companies.

The visualization system contains three modes. In the first mode, “Overall,” we plotted all the companies on a 2D plane. The color for each point is assigned based on the cluster to which it belongs. For example, in Exhibit 14a, we present all of the companies for the year 2020 and assign colors based on the ML model’s industry-level clusters. We can observe that General Motors Company and Ford Motor Company, both being automobile firms, belonged to the same industry group. Similarly, Dell Technologies (technology, hardware, storage), Visa (IT services), and Microsoft Corporation (software) were plotted close to each other, as they were learned to be similar by our model.

In the second mode, “Kernel(s),” we can select at least one company and plot the other companies belonging to the same cluster group on the 2D/3D plane. Exhibit 14b shows one such example for S&P Global’s ML sub-industry clustering for the year 2020. We can see that the NASDAQ Stock Market and Moody’s Investor Service were assigned to the same sub-industry group; they belong to the same GICS sub-industry group (financial exchange and data) group as S&P Global, whereas Cohen & Steers Inc., which belongs to “asset management and custody bank” in GICS, is also assigned to the same ML sub-industry.

In the third mode, “Nearest-Neighbor,” we can plot up to the 20 most-similar companies for a selected company X; these companies do not necessarily need to be classified in the same group as company X. Exhibit 14c shows the 10 nearest neighbors for FedEx Corporation (air freight and logistics). As expected, other transport companies like United Parcel Services (air freight and logistics) and CSX (railroads) were found to be similar; interestingly, H.B. Fuller Company (chemical), was also determined to be one of its 10 nearest neighbors (a chemical company according to GICS). These visualization frameworks can be accessed via https://dash-app-vis.herokuapp.com/.
(a) Overall mode. This figure shows all the companies for 2020 with a color assignment based on the industry group determined by the ML model. We observed that similar companies like Ford and General Motors belonged to the same industry clusters. Similar companies such as Microsoft, Visa, and Dell Technologies were placed close to each other as well.

(b) Kernel(s) mode. This figure shows the ML sub-industry cluster distribution for S&P Global for 2020.

(c) Nearest-neighbor mode. This figure shows the 10 nearest neighbors for FedEx in 2020.

Exhibit 14: Examples from the Cluster Visualization Framework.
Conclusion

In this paper, we developed an AI-driven peer grouping system, AIPGS, to group stocks according to market perception. These industry groupings are dynamic and updated as the characteristics of companies in the market evolve over time. As evidenced by its higher return correlations, AIPGS outperformed the closest benchmark, GICS, in grouping stocks. AIPGS uses information from multiple different data sources, including returns, factor exposure, 10-K filings, and news, to learn the similarity between companies, and uses that similarity to create groups of homogeneous companies.

AIPGS clusters had high intra-cluster similarity and high inter-cluster dispersion for financial ratios and fundamentals. We verified this for different fundamental characteristics including market beta, dividend yield, earnings yield, growth, residual volatility and size. These data-driven clusters grouped companies with similar fundamental characteristics, and we saw a higher level of diversification potential for portfolios constructed using ML clusters.

Apart from grouping stocks, our models predict a score for every pair of firms in our universe to reflect their future returns correlation (similarity). This can be used to obtain the nearest neighbors of a company, and to help investors understand which companies appear to be the most similar according to the market. In this work, we showed one use case of similarity scores: hedging single-stock portfolios.

This work showed that a classification system could be obtained using AI tools that performed better than traditional classification systems in some metrics, but the use of AI tools creates an interpretability problem. Although we can quantify the contributions of the datasets and various features for the similarity score of a pair of peer companies, we cannot assign a summary to the ML clusters that explains why the companies were grouped together. Thus, for some use cases where a premium is placed on interpretability and stability, traditional classification systems may be preferred. In other situations, where estimating future returns correlation is most important, our framework may dominate.

Regardless, we believe the framework presented here could be a valuable tool for investors, with potential applications for portfolio construction, hedging, and other areas. As an extension of this work, the ML clusters can be combined with the insights of experts and analysts to assign a summary to the clusters that can be interpreted by humans, as exists for GICS clusters. We also believe that an emphasis on greater interpretability may increase the accessibility of the ML clusters.

As we developed different versions (clusters from different ML methods) of AIPGS and compared them with GICS, we expect that different systems could be used, depending on the requirements of the user. We were able to obtain the best peers and clusters (in terms of forward returns correlation and similar risk–return profile) using the XGBoost model and sacrificing interpretability and stability of clusters. GICS had the most stable and interpretable clusters. Clusters and peers from the ridge regression model had a fair balance of stability and performance.
For investors, our peer grouping system, in conjunction with other existing classification systems, has the potential to be very insightful in investment diligence. For example, an investor looking for exposure to insurance brokers might include companies present in the GICS “insurance brokers” sub-industry. However, our model suggested including Berkshire Hathaway Inc. in this group. More specifically, for the year 2019, the ML model predicted that Berkshire Hathaway, a major conglomerate with investments across multiple industries, should be grouped in the same sub-industry as other firms that GICS would consider insurance brokers. On further analysis, this classification is meaningful because the market may perceive Berkshire Hathaway as an insurance broker, due to the fact that one of its most prominent companies, The Government Employees Insurance Company (GEICO), is an insurance underwriter. Because AIPGS provides additional insights beyond GICS, it can used as an extension of GICS. Furthermore, the similarity scores and clusters produced by our model may be used by some for portfolio management, such as investors seeking to improve hedging techniques.

Apart from the transparency and stability of ML-based clusters, another limitation of our approach is that our ML models can only group public companies owing to unavailability of data on private companies. Extending the classification system to non-public companies is one important extension that we hope to pursue in the future. Lastly, we can try to include other data sources, including analyst-coverage overlap, supply chain overlap, and broker research, that might contain extra information not captured in the datasets used in this work.
References

Albuquerque, A. (2009). Peer firms in relative performance evaluation. *Journal of Accounting and Economics* 48(1), 69–89.

Bayraktar, M., I. Mashtaler, N. Meng, and S. Radchenko (2015). Barra us total market equity trading model empirical notes. *MSCI Model Insight*.

Bhojraj, S., C. M. Lee, and D. K. Oler (2003). What’s my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41(5), 745–774.

Chan, L. K., J. Lakonishok, and B. Swaminathan (2007). Industry classifications and return comovement. *Financial Analysts Journal* 63(6), 56–70.

Dai, A. M., C. Olah, and Q. V. Le (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Dalziel, M., X. Yang, S. Breslav, A. Khan, and J. Luo (2018). Can we design an industry classification system that reflects industry architecture? *Journal of Enterprise Transformation* 8(1-2), 22–46.

De Bodt, E., B. E. Eckbo, and R. Roll (2020). Corporate rivalry and return comovement. *Tuck School of Business Working Paper* (3218544).

Fama, E. F. and K. R. French (1997). Industry costs of equity. *Journal of Financial Economics* 43(2), 153–193.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics* 116(1), 1–22.

Gay, S. and E. Karger (2014). Patent citations and stock performance: Constructing a dynamic industry classification. *Available at SSRN 2496414*.

Grover, A. and J. Leskovec (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864.

Harvey, C. R., Y. Liu, and H. Zhu (2015, 10). . . . and the Cross-Section of Expected Returns. *The Review of Financial Studies* 29(1), 5–68.

Heston, S. L. and K. G. Rouwenhorst (1995). Industry and country effects in international stock returns. *The Journal of Portfolio Management* 21(3), 53–58.

Hoberg, G. and G. Phillips (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124(5), 1423–1465.

Lee, C. M., P. Ma, and C. C. Wang (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116(2), 410–431.
Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance* 66(1), 35–65.

Lundberg, S. M., G. G. Erion, and S.-I. Lee (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* 11(1), 193–197.

Methodology, M. D. (2020). Global industry classification standard (GICS®) methodology: Guiding principles and methodology for gics.

MSCI (2021). MSCI peer similarity scores overview. *MSCI Inc.*.

Solnik, B. and J. Roulet (2000). Dispersion as cross-sectional correlation. *Financial Analysts Journal* 56(1), 54–61.

Sprenger, T. and I. M. Welpe (2011). Tweets and peers: Defining industry groups and strategic peers based on investor perceptions of stocks on twitter. *eBusiness & eCommerce eJournal*.

Van der Maaten, L. and G. Hinton (2008). Visualizing data using t-sne. *Journal of Machine Learning Research* 9(11).

Vermorken, M. A. (2011). Gics or icb, how different is similar? *Journal of Asset Management* 12(1), 30–44.

Zhang, H. and G. Bonne (2019). Peering into peer selection: Quantifying company similarity. *MSCI Research Insight*.

Zhang, H. and G. Bonne. (2020). Looking for a better hedge. *MSCI Research Insight*.
Data

Dataset Coverage

Exhibit A1 presents the different datasets and the number of companies in each dataset each year. The starting universe for different datasets is the MSCI USA Investable Market Index (IMI) universe. The index covers approximately 99% of the free-float-adjusted market capitalization in the U.S. For the models trained with all of the datasets, we took the inner join of the companies present in all datasets.

| Year | Returns-Daily | Returns-Monthly | Fundamental Data | 10-K Filings | News |
|------|---------------|------------------|------------------|--------------|------|
| 2011 | 2544          | 2542             | 2544             | 2203         | 2500 |
| 2012 | 2470          | 2470             | 2470             | 2227         | 2433 |
| 2013 | 2449          | 2449             | 2449             | 2189         | 2388 |
| 2014 | 2511          | 2510             | 2511             | 2238         | 2456 |
| 2015 | 2555          | 2552             | 2555             | 2324         | 2492 |
| 2016 | 2478          | 2477             | 2478             | 2288         | 2427 |
| 2017 | 2432          | 2431             | 2432             | 2255         | 2380 |
| 2018 | 2441          | 2441             | 2441             | 2271         | 2335 |
| 2019 | 2411          | 2411             | 2411             | 2196         | 2326 |
| 2020 | 2375          | 2372             | –                | –            | –    |

Exhibit A1: Datasets used for feature extraction and the number of companies covered in each dataset every year.

Factor Exposure Features

Table A1 presents the different factors used in our various models, along with their description. We also include the speed (slow or fast) with which the feature value changes. The slow factors were included at a yearly frequency, and fast factors were included at a monthly frequency.

| Factor Exposure Description | Speed of Change |
|-----------------------------|-----------------|
| One-day reversal            | Fast            |
| Historical Beta             | Slow            |
| Dividend-to-price ratio     | Slow            |
| Analyst-predicted dividend-to-price ratio | Slow |
| Lower partial moment        | Fast            |
Idiosyncratic lower partial moment  
Hybrid tail covariance risk  
Idiosyncratic hybrid tail covariance risk  
Mean lower partial moment CAPM beta  
Accruals - balance sheet version  
Accruals - cash flow statement version  
Variability in sales  
Variability in earnings  
Variability in cash flows  
Standard deviation of analyst prediction to price  
Enterprise Multiple (EBITDA to EV)  
Trailing Earnings-to-Price Ratio  
Analyst-predicted earnings-to-price ratio  
Long term analyst-predicted growth  
Historical earnings per share growth rate  
Historical sales per share growth rate  
Industry momentum  
Market Leverage  
Book Leverage  
Debt-to-Assets Ratio  
Monthly share turnover  
Quarterly share turnover  
Annual share turnover  
Modified Amihud illiquidity measure  
Pastor-Stambaugh illiquidity measure  
Long-term relative strength  
Long-term historical alpha  
Asset growth  
Issuance growth  
Capital expenditure growth  
Capital expenditure  
Cube of size exposure  
Relative strength  
Asset turnover  
Gross profitability  
Gross margin  
Return on assets  
Return on equity  
Skewness  
Maximum drawdown  
Regional momentum  
Historical Sigma  
Volatility implied by call options, 1 month  
Volatility implied by put options, 1 month  
Volatility implied by call options, 3 month
Volatility implied by put options, 3 month  Slow
Seasonality  Fast
Revision ratio  Fast
Change in analyst-predicted earnings-to-price  Fast
Change in analyst-predicted earnings per share  Fast
Positive sentiment based on composite sentiment Score  Fast
Positive sentiment based on event sentiment Score  Fast
Sentiment dispersion based on composite sentiment Score  Fast
At-the-money skew  Fast
Short interest  Fast
Short-term reversal  Fast
Log of market capitalization  Slow
Book-to-price ratio  Slow
Sales-to-price ratio  Slow
Cash flow to price ratio  Slow
Structural value  Slow

Table A1: Description of different factors and their speed of change.

**TF-IDF**

For featurization of the 10-K filing documents, all documents from 1994 to 2010 were used as training data (corpus). The n-grams of lengths 1 and 2 were considered to learn the vocabulary. Once the vocabulary was learned from all the documents, the documents from 2011 to 2019 were featurized for their use in the models. Each document was represented by an 82606-dimension sparse vector. Given the high dimensionality and sparse nature of TF-IDF features, we performed a truncated SVD to decrease the number of dimensions to 100. The implementation from sklearn and nltk libraries were used.

**Distance Metric**

In this section, we provide details about the cosine distance model and the ML models not included in the main text.

**Cosine Distance**

Given feature vectors $f_i^k$ and $f_j^k$, we determined the cosine distance. $i$ and $j$ indicate the company, and $k$ represents the dataset, such as past returns, 10-K filings, news, or factor exposure. The cosine distance ($d_{i,j}$) between two companies is calculated as $1 - \cos(f_i^k, f_j^k)$. When there are multiple datasets, the cosine distance is calculated for all datasets individually and then averaged. We believe that the cosine distance between the companies’ features is a strong predictor of the companies’ similarity because the cosine distance indicates the relative closeness between a pair of companies.

28
Machine Learning Models

Ridge Regression: This is a linear ML model widely used for regression tasks. It is a least squares model with an $L_2$ norm penalty on the weights that increases the generalization power of the model. We used the sklearn implementation of ridge regression. The linearity of the model makes its results interpretable and its input–output relationship transparent.

Neural Network: These are nonlinear supervised learning models used to learn the relationship between input features and the output. They consist of multiple layers that act as a computational engine. In neural networks, the values of the hidden units are computed as

$$h_j(x) = f(w_j + \sum_{i=0}^{n} w_{ij} \cdot x_i)$$

Here, $w_{ij}$ is the weight from input $x_i$ to the hidden unit, $h_j$. The weights, $w_{ij}$, can be learned by minimizing the loss function using optimizers like stochastic gradient descent.

The weights of the neural network create complex nonlinear interactions between input features, which makes the decision boundaries learned by neural networks nonlinear. The complex relationships learned in neural networks generally do not have human-interpretable meanings. Hence, the performance improvement obtained by using neural networks comes at the cost of model interpretability. In this work, we used neural networks for regression. We applied the implementation from tensorflow-keras for the construction and training of the model.

XGBoost: XGBoost is a gradient-boosted decision-trees model. A single decision tree is a supervised learning model that learns simple decision rules to segregate data points into different bins. The average of the outputs for all data points reaching the bin is considered as the ultimate output. Each tree is constructed using the classification and regression tree (CART) algorithm. Each tree acts a weak model. XGBoost is based on the idea of combining multiple weak models to create a strong model. Hence, it uses gradient descent to minimize loss while adding new models to the ensemble. Regression trees in general are interpretable models. It is easier to analyze the workings of a single regression tree using if-else conditions; however, XGBoost combines multiple decision trees to make the learned relationship nonlinear. The set of if-else conditions from multiple decision trees is intractable and difficult to interpret.

Dataset Split

Exhibit C1 gives the details of the training–validation–testing split for the different models trained for various years.
Exhibit C1: The Table above gives the train-validation-test year split. The testing year is an out-of-sample year. For example, when we learn the model for grouping companies for the year 2020, the input data (features) is from 2019. For validation, we use the input data from 2018 predicting groupings for 2019, and for training, we use the input data from 2017 predicting groupings for 2018.

| Test Year | Test Inputs | Valid. Year | Valid. Inputs | Train. Year | Train. Inputs |
|-----------|-------------|-------------|---------------|-------------|---------------|
| 2020      | 2019        | 2019        | 2018          | 2018        | 2017          |
| 2019      | 2018        | 2018        | 2017          | 2017        | 2016          |
| 2018      | 2017        | 2017        | 2016          | 2016        | 2015          |
| 2017      | 2016        | 2016        | 2015          | 2015        | 2014          |
| 2016      | 2015        | 2015        | 2014          | 2014        | 2013          |
| 2015      | 2014        | 2014        | 2013          | 2013        | 2012          |
| 2014      | 2013        | 2013        | 2012          | 2012        | 2011          |

Model and Results

Cosine Model Evaluation

Exhibit D1 reports the average return correlations for the peers obtained from the cosine model. The first row corresponds to the GICS model, and the other rows correspond to cosine models that are based on different datasets. The cosine model performed best when all the datasets were combined, as represented by the row “ALL.” However, when compared to GICS, the improvements were minor. Across different datasets, the cosine distance of past returns was the best predictor of future returns correlations, followed by different variants of 10-K features, factors, and news.

After we obtained the cosine similarity, we grouped the companies using hierarchical clustering. Exhibit D2 presents the average pairwise correlation for clusters obtained using hierarchical clustering on cosine distances. The number of peers (\( k \) in top \( k \)) is decided based on the number of companies belonging to the cluster at the specified level. The performance of the clusters was comparable to GICS clusters (except for sub-industry groupings). This implies the need for more complex models, such as ML models, to predict the similarity between companies.

Performance by Year

In Exhibit D3, we present the average pairwise return correlation between peer companies for the ML model. For a fair comparison with GICS, the number of peers (\( k \) in top \( k \)) was decided based on the number of companies belonging to the GICS group of the company at the specified sector, industry, or sub-industry level. The ML models consistently outperformed
Exhibit D1: Evaluation of Cosine Model. The table presents the average return correlations for peers in the out-of-sample period in which the peers are decided via the cosine model for different datasets. For a fair comparison with GICS, the number of peers for sector, industry, and sub-industry groupings was decided according to the number of companies in the GICS group to which the company belongs. We observed that the performance improvements using all datasets combined (ALL) was small compared to GICS.

| Features          | Sector | Industry | Sub-industry | Top 10 | Top 5 | Top 2 |
|-------------------|--------|----------|--------------|--------|-------|-------|
| GICS              | 32.40  | 41.23    | 43.79        | -      | -     | -     |
| News              | 27.67  | 30.15    | 31.79        | 33.44  | 35.23 | 37.97 |
| Factors           | 31.45  | 37.61    | 39.99        | 39.80  | 41.54 | 44.34 |
| 10K-TF-IDF(SVD)   | 29.87  | 37.99    | 42.26        | 40.97  | 43.00 | 45.71 |
| 10K-TF-IDF        | 31.66  | 40.10    | 43.09        | 40.88  | 42.91 | 45.76 |
| 10K-ALL           | 31.11  | 39.49    | 43.12        | 41.51  | 43.49 | 46.14 |
| Returns-daily     | 32.56  | 41.12    | 44.00        | 43.76  | 45.86 | 48.96 |
| ALL               | 32.55  | **41.42**| **44.34**    | **44.15** | **46.55** | **49.97** |

Exhibit D2: Results after hierarchical clustering: Cosine Model. This table presents the average return correlations for the clusters obtained using the cosine model. The number of peers at each level was determined using the number of companies belonging to the cluster at that level of granularity. The cosine model clusters failed to outperform GICS at the sector and industry levels, implying the need of a more complex model, such as ML modeling, to obtain the relative distances between companies.

| Features          | Sector | Industry | Sub-industry |
|-------------------|--------|----------|--------------|
| GICS              | **32.40** | **41.23** | 43.79        |
| News              | 26.55  | 29.40    | 31.39        |
| Factors           | 30.32  | 36.33    | 38.70        |
| 10K-TF-IDF(SVD)   | 26.04  | 36.99    | 41.55        |
| 10K-TF-IDF        | 30.69  | 39.18    | 43.40        |
| 10K-ALL           | 30.62  | 39.77    | 42.11        |
| Returns-daily     | 28.69  | 38.04    | **45.46**    |
| ALL               | 31.40  | 39.96    | 44.05        |
GICS over all years. The best-performing models are highlighted in bold. The XGBoost model trained on all features, including GICS, was the best-performing model over the years at various levels. We also include a column for the maximum possible average correlation values, calculated by selecting peers on the basis of the forward returns correlation.

In Exhibit D4, from 2014–2020, we present the average return correlation across all clusters obtained from the different ML models from 2014–2020. The number of peers \((k\) in top \(k\)) was decided based on the number of companies belonging to the cluster of the company. The best-performing models over the years are highlighted in bold. The clusters using ML models outperformed the GICS clusters by a large margin at the industry and sub-industry levels.

**Bootstrapping**

In Exhibit D3 we observed that the ML models outperformed GICS clusters consistently. We obtained an increase in the forward return correlation up to 4%. To analyze the statistical significance of the improvements, we performed data bootstrapping, and trained a ridge regression (Ridge+GICS) model and calculated the forward returns correlation for the peers predicted by the model. We repeated the process 100 times and obtained the standard errors of the forward returns correlation values for the peers predicted by the model (Exhibit D5). We observed that errors were much smaller (by an order of magnitude) than the performance improvement obtained. We did not perform bootstrapping for the neural network and XGBoost models owing to the large training time of those models.

**Cluster Size**

The aim of hierarchical clustering is to avoid having a majority of companies assigned to very few clusters. To evaluate this, we computed the standard deviation of cluster size at the sector, industry, and sub-industry levels, because uniformly distributed clusters will have smaller standard deviations. Exhibit D6 presents the standard deviation of the cluster-size distribution. It shows that companies are more uniformly distributed in the ML clusters compared to GICS for industry and sub-industry groupings.

**Dataset Importance**

Exhibit D7 presents the importance of different datasets over time. Red indicates higher values, and blue indicates lower ones. From 2014 to 2020, the returns were consistently the most important dataset.
| Year | Max | GICS | Ridge | Ridge+GICS | NN | NN+GICS | XGB | XGB+GICS |
|------|-----|------|-------|------------|----|---------|-----|----------|
|      |     |      |       |            |    |         |     |          |
| Sector |
| 2014 | 40.33 | 30.17 | 34.06 | **34.10** | 33.57 | 33.81 | 34.02 | 33.78 |
| 2015 | 42.70 | 32.43 | 36.45 | 36.47 | 36.37 | 36.82 | 36.97 | **37.08** |
| 2016 | 43.26 | 33.91 | 36.91 | 36.97 | 36.97 | 36.85 | 37.20 | **37.30** |
| 2017 | 30.92 | 22.51 | 23.38 | 23.25 | 23.34 | 23.85 | 23.75 | **23.92** |
| 2018 | 42.85 | 33.66 | 35.79 | 35.88 | 36.02 | 36.09 | 35.87 | **36.42** |
| 2019 | 40.15 | 29.79 | 33.29 | 33.24 | 32.97 | 32.94 | 33.45 | **33.38** |
| 2020 | 61.87 | 49.76 | 53.51 | 53.54 | 53.68 | 53.76 | **55.10** | 54.09 |
| Industry |
| 2014 | 48.39 | 38.33 | 40.77 | **40.93** | 39.76 | 40.01 | 40.40 | 40.29 |
| 2015 | 52.37 | 42.79 | 45.05 | 45.30 | 44.91 | 45.43 | 45.55 | **45.79** |
| 2016 | 51.97 | 42.99 | 44.82 | 45.00 | 45.01 | 44.82 | 45.33 | **45.55** |
| 2017 | 40.86 | 32.08 | 32.87 | 33.20 | 32.43 | 33.29 | 33.09 | **33.45** |
| 2018 | 51.73 | 43.06 | 44.03 | 44.47 | 44.37 | 44.36 | 44.03 | **45.01** |
| 2019 | 47.71 | 38.07 | 39.86 | 40.01 | 40.05 | 40.04 | 40.30 | **40.42** |
| 2020 | 67.68 | 56.44 | 58.82 | 58.83 | 58.92 | 59.13 | **59.38** | 59.33 |
| Sub-industry |
| 2014 | 51.90 | 41.53 | 43.92 | **44.10** | 42.70 | 42.91 | 43.66 | 43.57 |
| 2015 | 54.39 | 44.47 | 46.77 | 47.06 | 46.75 | 47.14 | 47.10 | **47.39** |
| 2016 | 54.32 | 45.69 | 47.30 | 47.55 | 47.33 | 47.12 | 47.72 | **47.90** |
| 2017 | 44.61 | 36.08 | 36.80 | **37.25** | 36.39 | 37.19 | 36.92 | 37.23 |
| 2018 | 54.14 | 45.28 | 46.37 | 46.81 | 46.89 | 46.86 | 46.70 | **47.30** |
| 2019 | 50.19 | 41.03 | 42.40 | 42.55 | 42.66 | 42.70 | 42.75 | **42.97** |
| 2020 | 67.95 | 57.93 | 59.97 | 60.07 | 59.94 | 60.23 | 60.48 | **60.50** |

Exhibit D3: This table presents the pairwise return correlation over years when the number of peers (top k) is selected by considering the number of peers in the corresponding GICS cluster (for fair comparison with GICS). The best model for each year is bolded.
| Year | GICS   | Ridge | Ridge+GICS | NN    | NN+GICS | XGB   | XGB+GICS |
|------|--------|-------|------------|-------|---------|-------|----------|
|      |        |       |            |       |         |       |          |
|      |        |       |            |       |         |       |          |
|      |        |       |            |       |         |       |          |
| Year | Sector |       |            |       |         |       |          |
| 2014 | 2014   | 30.17 | 30.49      | 30.57 | 28.35   | 30.12 | 31.13    |
| 2015 | 2015   | 32.43 | 32.27      | 30.70 | 29.42   | 30.78 | 31.86    | 32.69    |
| 2016 | 2016   | 33.91 | 32.15      | 34.34 | 31.48   | 31.87 | 33.56    | 33.51    |
| 2017 | 2017   | 22.51 | 19.60      | 22.04 | 20.65   | 20.41 | 19.24    | 21.01    |
| 2018 | 2018   | 33.66 | 30.79      | 34.88 | 30.25   | 29.05 | 32.57    | 34.66    |
| 2019 | 2019   | 29.79 | 30.11      | 30.63 | 27.04   | 24.12 | 28.10    | 28.78    |
| 2020 | 2020   | 49.76 | 48.89      | 51.63 | 47.67   | 49.82 | 50.22    | 48.52    |

| Year | Industry |       |            |       |         |       |          |
|------|----------|-------|------------|-------|---------|-------|----------|
| 2014 | 2014     | 38.33 | 41.12      | 41.24 | 31.97   | 34.54 | 39.47    | 41.44    |
| 2015 | 2015     | 42.79 | 43.70      | 45.04 | 41.12   | 41.40 | 40.77    | 44.53    |
| 2016 | 2016     | 42.99 | 42.95      | 44.51 | 41.10   | 41.69 | 44.36    | 45.65    |
| 2017 | 2017     | 32.08 | 38.84      | 36.26 | 33.83   | 34.81 | 34.80    | 34.74    |
| 2018 | 2018     | 43.06 | 45.11      | 45.10 | 42.41   | 44.32 | 43.98    | 45.07    |
| 2019 | 2019     | 38.07 | 40.85      | 41.87 | 38.40   | 40.30 | 38.78    | 40.67    |
| 2020 | 2020     | 56.44 | 60.14      | 59.08 | 58.83   | 60.86 | 59.30    | 60.73    |

| Year | Sub-industry |       |            |       |         |       |          |
|------|--------------|-------|------------|-------|---------|-------|----------|
| 2014 | 2014        | 41.53 | 45.36      | 45.74 | 40.23   | 39.98 | 45.57    | 45.14    |
| 2015 | 2015        | 44.47 | 47.87      | 49.45 | 46.27   | 48.74 | 49.11    | 50.93    |
| 2016 | 2016        | 45.69 | 48.60      | 50.13 | 49.31   | 48.18 | 49.35    | 50.21    |
| 2017 | 2017        | 36.08 | 41.47      | 42.48 | 37.92   | 39.51 | 38.62    | 42.61    |
| 2018 | 2018        | 45.28 | 48.61      | 51.27 | 50.91   | 50.63 | 48.12    | 51.86    |
| 2019 | 2019        | 41.03 | 45.35      | 47.29 | 47.61   | 46.73 | 48.63    | 47.82    |
| 2020 | 2020        | 57.93 | 63.95      | 64.11 | 65.15   | 66.00 | 66.76    | 65.69    |

Exhibit D4: This table presents the pairwise return correlation from 2014 to 2020 for GICS clusters and different ML model clusters. The number of peers (k in top k) is decided by the number of companies belonging to the cluster of the company at a specified level. The best model for each year is bolded. For industry and sub-industry clusters, the average pairwise return correlation is higher than for GICS clusters by 10-20%.
| Year | Sector | Industry | Sub-industry | Top-2 | Top-5 | Top-10 |
|------|--------|----------|--------------|-------|-------|--------|
| 2014 | 0.0057 | 0.0078   | 0.0097       | 0.0439| 0.0208| 0.0157 |
| 2015 | 0.0104 | 0.0151   | 0.0160       | 0.0367| 0.0240| 0.0206 |
| 2016 | 0.0055 | 0.0064   | 0.0077       | 0.0485| 0.0239| 0.0163 |
| 2017 | 0.0091 | 0.0105   | 0.0110       | 0.0646| 0.00369| 0.0286 |
| 2018 | 0.0062 | 0.0083   | 0.0080       | 0.0433| 0.0218| 0.0159 |
| 2019 | 0.0082 | 0.0069   | 0.0061       | 0.0345| 0.0183| 0.0149 |
| 2020 | 0.0067 | 0.0075   | 0.0077       | 0.0400| 0.0199| 0.0144 |

Exhibit D5: This table presents the standard errors of the correlation in the returns for the peers obtained from the ridge regression model. The standard errors were smaller by orders of magnitude when compared to the improvements obtained. The reported values are in percentages.

| Model:Features          | Sector | Industry | Sub-industry |
|-------------------------|--------|----------|--------------|
| GICS                    |        |          |              |
| Ridge:ALL               | 124.10 | 29.65    | 16.41        |
| Ridge:ALL+GICS          |        |          |              |
| NN:ALL                  | 188.95 | 27.98    | 15.15        |
| NN:ALL+GICS             |        |          |              |
| XGBoost:ALL             | 140.24 | 26.54    | 14.89        |
| XGBoost:ALL+GICS        | 159.92 | 27.34    | 13.79        |
| XGBoost:ALL+GICS        | 161.35 | 26.32    | 14.47        |
| XGBoost:ALL+GICS        | 124.10 | 29.65    | 16.42        |
| XGBoost:ALL+GICS        | 133.50 | 23.37    | 13.50        |

Exhibit D6: The table presents the standard deviation of cluster sizes averaged over years. The clusters obtained using ML models were more uniformly distributed than GICS clusters at the industry and sub-industry levels.

|            | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|------------|------|------|------|------|------|------|------|
| GICS       | 0.05 | 0.07 | 0.08 | 0.09 | 0.10 | 0.14 | 0.10 |
| News       | 0.12 | 0.12 | 0.12 | 0.12 | 0.13 | 0.15 | 0.15 |
| 10k-D2V    | 0.18 | 0.18 | 0.19 | 0.19 | 0.20 | 0.17 | 0.19 |
| 10k-TFIDF  | 0.19 | 0.18 | 0.22 | 0.18 | 0.22 | 0.20 | 0.18 |
| Factors    | 0.24 | 0.36 | 0.22 | 0.25 | 0.26 | 0.31 | 0.37 |
| Returns    | 0.43 | 0.39 | 0.39 | 0.46 | 0.45 | 0.41 | 0.45 |

Exhibit D7: The above figure presents the importance of features from different datasets aggregated. The returns were consistently the most important dataset from 2014 to 2020.