HiveCluster: Strategic Stock Portfolio Construction

Lucas O. Martínez
Department of Computer Science
New York University
New York City, USA
lom2017@nyu.edu

Prasanna A P
Department of Computer Science
New York University
New York City, USA
pa2490@nyu.edu

***Abstract*—In the fast-paced world of finance, identifying interdependent stocks within major indices like the Standard & Poor's 500 (S&P 500) is crucial for optimizing investment strategies and portfolio management. Systems that predict groups of related stocks and construct portfolios based on user-defined risk tolerance and investment periods offer investors a competitive edge, enhancing decision-making and risk assessment. This paper presents *HiveCluster*, a system that identifies stock relationships and clusters within the S&P 500 to improve portfolio construction aligned with individual risk profiles. After experimenting with multiple algorithms and datasets, we focused on clustering daily percentage changes in historical stock prices to obtain more meaningful groupings. We incorporate rolling metrics such as volatility, Sharpe ratio, beta, and maximum drawdown to classify these clusters into four investor risk tolerance categories: conservative, moderate, moderately aggressive, and aggressive. Empirical results show that *HiveCluster* consistently forms cohesive clusters and outperforms the S&P 500 benchmark over short testing periods, demonstrating the potential of data-driven clustering and risk-based portfolio construction to enhance long-term returns.**

Keywords— data science, predictive analytics, machine learning, stock market, market leaders, portfolio, hedge fund, lead-lag relationship, clustering

I. Introduction

In an evolving financial landscape, the ability to identify and exploit relationships between stocks is a cornerstone of effective portfolio management. The Standard & Poor's 500 (S&P 500) index, comprising 500 leading U.S. companies, exhibits complex interdependencies among its constituents. Understanding these relationships allows investors to anticipate market movements, manage risk, and enhance returns.

This paper introduces a data-driven framework, *HiveCluster*, that clusters S&P 500 stocks based on their historical performance. We then refine and categorize these clusters according to different risk tolerance levels. The approach is conducted in two key phases:

a) **Clustering S&P 500 Stocks into Buckets**: We apply unsupervised learning (K-means, DBSCAN, Agglomerative Clustering, and others) to daily percentage changes in stock prices. Additionally, we performed exploratory tests on SEC Form 13F filings but found historical price data alone offered clearer, more robust clusters.

b) **Clustering Buckets into Risk Tolerance Categories**: By calculating rolling performance metrics (volatility, Sharpe ratio, beta, and maximum drawdown), we segment the stock clusters into four categories: *conservative, moderate, moderately aggressive, and aggressive*.

The proposed methodology aligns portfolio construction with individual risk tolerance and investment horizons. Empirical results demonstrate promising performance, with clustered portfolios consistently outperforming the S&P 500 benchmark during our testing periods, especially in more aggressive risk profiles. The remainder of this paper details the relevant research, data sources, methodology, experiments, and results, followed by our conclusion and plans for future research.

II. Literature Review

Numerous studies have examined stock correlation and institutional investor behavior for asset allocation and trading strategies.

A. Stock Relationship Analysis

Miori and Cucuringu [1] leveraged SEC Form 13F data to compute trading imbalances and guide contrarian trading strategies. Although we explored leveraging 13F filings, we primarily rely on historical price data. Han et al. [2] introduced *Dynamic Time Warping* (DTW) to detect lead-lag relationships in stock prices, demonstrating that temporal alignment reveals interdependencies valuable for cluster analysis.

Gupta et al. [3] and Agarwal et al. [4] applied PCA-based dimensionality reduction and unsupervised learning (e.g., DBSCAN, K-means) for improving pairs trading performance and highlighting the importance of robust clustering. These frameworks inform our focus on correlation-based clustering for identifying cohesive stock groupings.

B. Sentiment Analysis in Finance

Sentiment analysis methods, such as those by Hung et al. [5], and Chen [6], have demonstrated that integrating financial news sentiment can enhance portfolio returns. While *HiveCluster* primarily uses quantitative data (prices, returns, risk metrics), future iterations could incorporate sentiment analysis to detect events that drive stock co-movements.

C. SEC Filings Analysis

Anderson and Brockman [7] highlighted inaccuracies in 13F filings, cautioning the overreliance on these reports. Seth et al. [8] developed a framework leveraging structured and unstructured data to uncover co-movement driven by co-ownership among institutional investors. Angelini et al. [9] pointed out the impact of concentrated hedge fund "top picks" on alpha generation. Although these insights are valuable, our experiments ultimately found that focusing on historical price data provided clearer, more robust clusters.

III. Datasets

A. Data Collection

Historical Stock Data: We acquired daily, weekly, and monthly historical data from the Yahoo Finance API for all S&P 500 constituents. The metrics included open, high, low, close, adjusted close, and trading volume.

SEC Form 13F: We scraped data from the SEC EDGAR database for the top 100 investment firms (*by AUM*) over a 10-year span (2014–2024). Although initially integrated into our clustering, 13F-based clusters were outperformed by purely price-driven clusters in most trials.

Company Details: Sector, industry, and general company information were collected from Yahoo Finance and WhaleWisdom.com to enrich post-clustering analysis.

B. Data Overview

a) Historical Stock Data: includes daily returns, necessary for correlation analysis, leading-lag detection, and clustering.

b) SEC Form 13F Data: logs institutional investors' quarterly holdings. We standardized CUSIP identifiers to stock tickers using the OpenFigi API.

c) S&P 500 Benchmark: daily returns were used to compute rolling metrics and compare against our resulting portfolios.

IV. Methodology

Figure 1 summarizes our pipeline:

1) **Data Extraction**
2) **Clustering Stocks into Buckets**
3) **Clustering Buckets into Risk Tolerance Categories**
4) **Portfolio Construction and Evaluation**

A. Data Extraction (and Engineering)

Our data extraction process consolidated diverse datasets to establish a robust foundation for analysis. Historical stock prices for S&P 500 companies were collected via the Yahoo Finance API, capturing *daily closing prices*, which served as the basis for correlation analysis and clustering. SEC Form 13F filings were scraped from the SEC EDGAR database, focusing on quarterly holdings of the top 100 investment firms ranked by assets under management. Stock identifiers were standardized using the OpenFigi API to ensure consistency across datasets. Company-specific details, such as sector and industry classifications, were also retrieved for integration.

Institutional Metrics: To analyze institutional investor behavior, two key metrics were derived from SEC Form 13F filings:

- *Average Term Conviction:* This metric measures hedge fund confidence in a specific stock during a reporting term:

$$\text{Avg Conviction}_{\text{term}}(T) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{Shares of ticker } T \text{ held by fund } i}{\text{Total shares held by fund } i}\right) \tag{1}$$

  Where N is the total number of hedge funds evaluated. Higher conviction values indicate stronger confidence in the stock.

- *Term Consensus:* This metric reflects the popularity of a stock among hedge funds during a reporting term:

$$\text{Consensus}_{\text{term}}(T) = \frac{\text{Num of funds holding ticker } T \text{ during the term}}{\text{Total num of funds reporting during the term}} \tag{2}$$

  Higher consensus values highlight broader popularity and widespread adoption of the stock.

Stock Relationships: To uncover relationships between stocks, we applied a two-step approach combining Granger causality and Pearson correlation:

- *Granger Causality:* Identified whether the past values of one stock could predict the future values of another. If *p-value* of Granger < 0.05, then causality exists.
- *Pearson Correlation:* Measured the strength and direction of these identified relationships.

The combined relationship was represented as:

$$\text{Coefficient}_{A,B} = I(\text{Granger}(A,B))\cdot\text{Sign}(Pearson(A,B)) \tag{3}$$

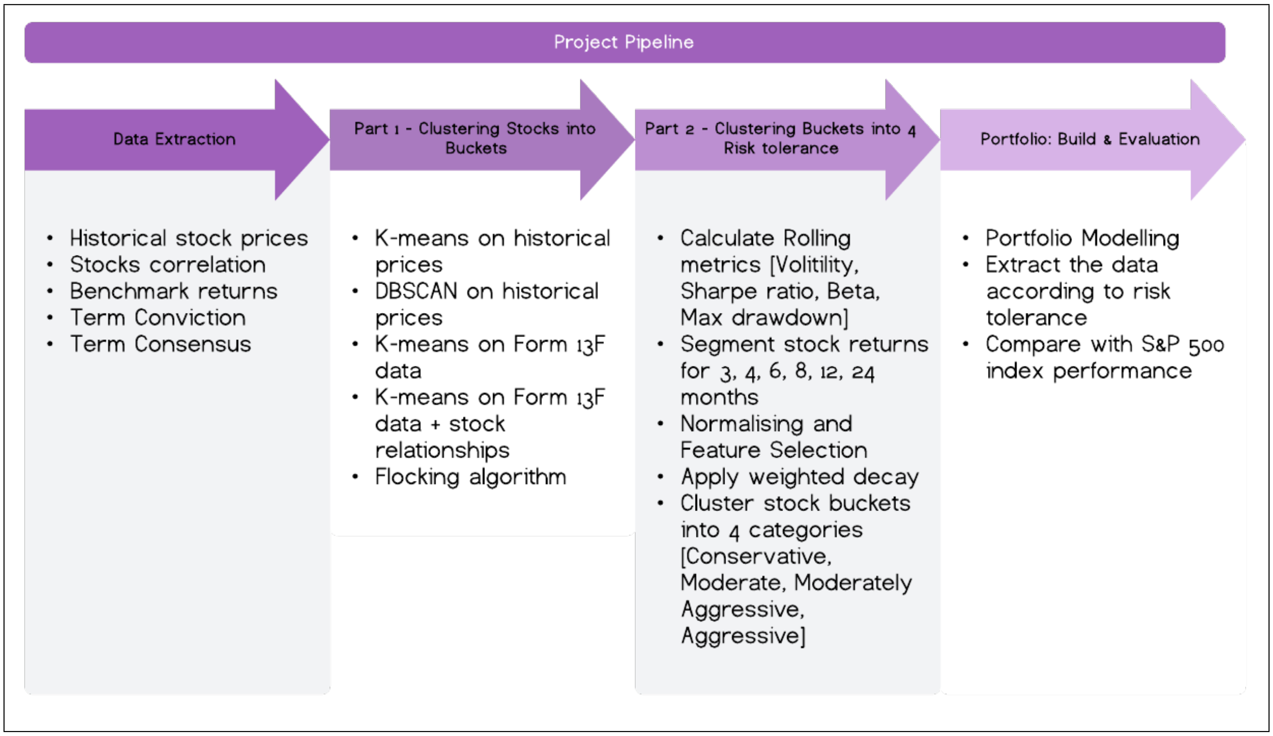

Project Pipeline
Data Extraction
Part 1 - Clustering Stocks into Buckets
Part 2 - Clustering Buckets into 4 Risk tolerance
Portfolio: Build & Evaluation
• Historical stock prices
• Stocks correlation
• Benchmark returns
• Term Conviction
• Term Consensus
• K-means on historical prices
• DBSCAN on historical prices
• K-means on Form 13F data
• K-means on Form 13F data + stock relationships
• Flocking algorithm
• Calculate Rolling metrics [Volitility, Sharpe ratio, Beta, Max drawdown]
• Segment stock returns for 3, 4, 6, 8, 12, 24 months
• Normalising and Feature Selection
• Apply weighted decay
• Cluster stock buckets into 4 categories [Conservative, Moderate, Moderately Aggressive, Aggressive]
• Portfolio Modelling
• Extract the data according to risk tolerance
• Compare with S&P 500 index performance

Fig. 1: Overview of the project pipeline.

where:

- $I(\text{Granger}(A, B))$: Indicator function, equal to 1 if Granger causality exists, otherwise 0.
- $\text{Sign}(Pearson(A, B))$: Direction of the Pearson correlation ($+1$ for positive, -1 for negative).

These metrics and coefficients provided critical inputs for clustering, allowing us to quantify both institutional behaviors and stock interdependencies.

B. Clustering Stocks into Buckets

We tested multiple algorithms: **K-means**, DBSCAN, and hierarchical (Agglomerative). We also experimented with a "Bird Flock Algorithm" approach inspired by local particle interactions. We used:

- **Daily percentage changes**: Derived from historical prices.
- **K-means on 13F Data**: Clustering by co-ownership patterns across top investment firms.
- **K-means using Price Data + Granger Causality + Pearson Correlation**: Enhanced features to detect directional lead-lag relationships.

Ultimately, **K-means on daily percent changes** in stock prices showed consistently better cluster cohesion and interpretability. We used metrics such as the *Silhouette Score* and a custom *Cluster Cohesion Score (CCS)* to guide the optimal choice of the number of clusters.

a) Cluster Cohesion Score (CCS): was used to ensure certain pairs of known highly related stocks (e.g., large banks, big tech) frequently ended up in the same clusters. For instance, Apple & Microsoft were expected to appear together, given known co-movement.

C. Clustering Buckets into Risk Tolerance Categories

Once the buckets were formed, we computed rolling metrics for each cluster:

i) **Volatility**: Standard deviation of daily returns
ii) **Sharpe Ratio**: Risk-adjusted returns
iii) **Beta**: Relative volatility vs. S&P 500
iv) **Maximum Drawdown**: Largest observed peak-to-trough decline

These metrics were averaged (or otherwise aggregated) within each cluster across multiple horizons (3, 4, 6, 8, 12, 24 months). To emphasize more recent performance, we applied a weighted decay factor:

$$W = \frac{\sum_{i=0}^{n-1} (v_i \cdot d^i)}{\sum_{i=0}^{n-1} d^i}, \tag{4}$$

where v_i is the value at time i, d is the decay rate ($0 < d \leq 1$), and n is the number of observations.

We then used **K-means** on these aggregated rolling metrics to classify clusters into four risk categories: *conservative, moderate, moderately aggressive, and aggressive*.

D. Portfolio Construction and Evaluation

To evaluate the performance of our framework, we constructed portfolios based on clusters categorized into different risk tolerance levels. Then, the model was tested over a 10-year training period (October 31, 2014 October 30, 2024) and a testing period for November, 2024, with a focus on a moderate risk tolerance profile. The evaluation process includes both portfolio construction and performance comparison against the S&P 500 benchmark.

V. Experiments and Results

To evaluate portfolio performance against the S&P 500 benchmark, we will randomly select four groups of stocks, with each group containing five stocks. These selected groups will reflect a diverse mix of sectors and risk profiles, aligning with the clustering results. The performance of these groups will be tracked over the same investment period (November, 2024) and directly compared to the S&P 500. This approach aims to assess how effectively the clustered portfolios perform relative to the broader market, providing insight into the model's accuracy and potential for generating competitive returns.

A. Clustering Algorithm Comparison

Table I summarizes major experiments conducted in *Stage 1: Clustering Stocks into Buckets*. We vary the training data sources and clustering algorithms. The *Cluster Cohesion Score (CCS)* scale is from 1–10, higher meaning more robust co-movement among known stock pairs.

TABLE I: Comparison of Clustering Algorithms and Data Inputs

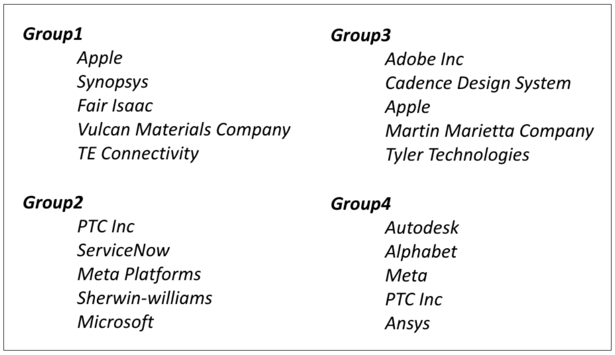
Training Data	Clustering Algorithm	CCS (1-10)	# Clusters
S&P 500 - daily % returns	K-means	9	117
S&P 500 - daily % returns	DBSCAN	6	129
S&P 500 - daily % returns	Agglomerative	4	78
S&P 500 - financial quarters	K-means	7	117
SEC Form 13F filings	K-means	4	75
SEC Form 13F + Granger + Pearson	K-means	7	150

The best-performing model is **K-means on daily % returns**, with a CCS of 9, producing 117 clusters. This method captures meaningful co-movements more consistently than the others.

B. Risk-Based Clustering and Portfolio Construction

After we form buckets (117 clusters), we roll up the performance metrics (volatility, Sharpe, beta, and maximum draw-down) over horizons (3–24 months) and apply weighted decay. We then segment the clusters into four risk categories using K-means on these aggregated statistics.

Figure 2 shows sample output. Each cluster is assigned to *Conservative, Moderate, Moderately Aggressive, or Aggressive*, according to the aggregated rolling metrics. Users then query the model for a specific risk tolerance, timeframe, and optionally filter by sector.



Group1
Apple
Synopsys
Fair Isaac
Vulcan Materials Company
TE Connectivity

Group2
PTC Inc
ServiceNow
Meta Platforms
Sherwin-williams
Microsoft

Group3
Adobe Inc
Cadence Design System
Apple
Martin Marietta Company
Tyler Technologies

Group4
Autodesk
Alphabet
Meta
PTC Inc
Ansys

Fig. 2: Example cluster categorization into four risk profiles for a 3-month horizon.

C. Performance Evaluation

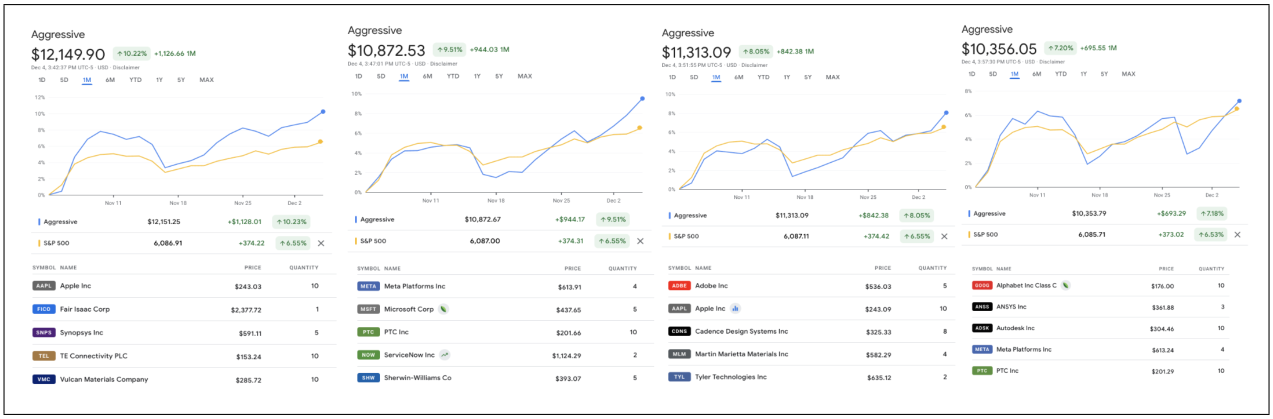
In final testing, we formed portfolios based on the **Aggressive** cluster selection over a 3-month horizon. Figure 3 illustrates four random 5-stock groups from the recommended *Aggressive* cluster.

All four sample portfolios outperformed the S&P 500 benchmark in short-term returns. For moderate risk profiles, the portfolio also delivered superior returns vs. the benchmark, supporting that these clusters capture meaningful co-movements and risk patterns.

VI. Conclusion

We introduced *HiveCluster*, a framework for identifying interdependent stocks in the S&P 500 and constructing risk-aware portfolios. Experimental results demonstrated that **K-means on daily price returns** outperforms other approaches for clustering. When combined with rolling performance metrics, these buckets offer tailored investment strategies for four distinct risk tolerance levels.

Our research contributes to an emerging body of work exploring data-driven, dynamic portfolio construction. While the results are promising, further stress testing (over longer horizons and different market conditions) will be conducted to enhance scalability. Future extensions include integrating *sentiment analysis* from financial news, exploring advanced optimization methods (e.g., Black-Litterman), and refining risk factor models. Unexpected cluster pairings also merit deeper investigation; for instance, discovering that two seemingly unrelated firms (e.g., Airbnb and Dr Pepper) cluster together might reveal subtle but impactful market influences or co-ownership patterns.



Aggressive

$12,149.90 ↑10.22% +1,126.66 1M

Dec 4, 3:42:37 PM UTC-5 · USD · Disclaimer

Aggressive	$12,151.25	+$1,128.01	↑10.23%
S&P 500	6,086.91	+374.22	↑6.55%

SYMBOL	NAME	PRICE	QUANTITY
AAPL	Apple Inc	$243.03	10
FICO	Fair Isaac Corp	$2,377.72	1
SNPS	Synopsys Inc	$591.11	5
TEL	TE Connectivity PLC	$153.24	10
VMC	Vulcan Materials Company	$285.72	10

Aggressive

$10,872.53 ↑9.51% +944.03 1M

Dec 4, 3:47:01 PM UTC-5 · USD · Disclaimer

Aggressive	$10,872.67	+$944.17	↑9.51%
S&P 500	6,087.00	+374.31	↑6.55%

SYMBOL	NAME	PRICE	QUANTITY
META	Meta Platforms Inc	$613.91	4
MSFT	Microsoft Corp	$437.65	5
PTC	PTC Inc	$201.66	10
NOW	ServiceNow Inc	$1,124.29	2
SHW	Sherwin-Williams Co	$393.07	5

Aggressive

$11,313.09 ↑8.05% +842.38 1M

Dec 4, 3:51:55 PM UTC-5 · USD · Disclaimer

Aggressive	$11,313.09	+$842.38	↑8.05%
S&P 500	6,087.11	+374.42	↑6.55%

SYMBOL	NAME	PRICE	QUANTITY
ADBE	Adobe Inc	$536.03	5
AAPL	Apple Inc	$243.09	10
CDNS	Cadence Design Systems Inc	$325.33	8
MLM	Martin Marietta Materials Inc	$582.29	4
TYL	Tyler Technologies Inc	$635.12	2

Aggressive

$10,356.05 ↑7.20% +695.55 1M

Dec 4, 3:57:30 PM UTC-5 · USD · Disclaimer

Aggressive	$10,353.79	+$693.29	↑7.18%
S&P 500	6,085.71	+373.02	↑6.53%

SYMBOL	NAME	PRICE	QUANTITY
GOOG	Alphabet Inc Class C	$176.00	10
ANSS	ANSYS Inc	$361.88	3
ADSK	Autodesk Inc	$304.46	10
META	Meta Platforms Inc	$613.24	4
PTC	PTC Inc	$201.29	10

Fig. 3: Performance of four random 5-stock groups compared to the S&P 500 benchmark.

VII. Future Work

Potential extensions to *HiveCluster* include:

a) **Feature Exploration**: Introducing market capitalization, advanced textual features (NLP on company reports).

b) **Sentiment Integration**: Analyzing financial news, social media, and crowd-sourced sentiment to detect event-driven co-movements in real time.

c) **Portfolio Stress Testing**: Running long-term simulations in virtual trading environments under various volatility regimes.

d) **Advanced Optimization Models**: Incorporating the Black-Litterman framework to fuse subjective investor outlook with data-driven signals.

Such directions aim to improve the system's robustness, adaptability, and overall utility in real-world portfolio construction and management.

References

[1] D. Miori and M. Cucuringu, "SEC Form 13F-HR: Statistical investigation of trading imbalances and profitability analysis," *arXiv*, 2022.

[2] T. Han, et al., "Discovering the lead-lag relationships in financial markets: A method based on DTW," in *2019 Chinese Automation Congress (CAC)*, IEEE, 2019, pp. 4262–4267.

[3] V. Gupta, V. Kumar, Y. Yuvraj, and M. Kumar, "Optimized pair trading strategy using unsupervised machine learning," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, IEEE, 2023, pp.1–5.

[4] A. Agarwal, et al., "Comparative study on pairs trading using machine learning algorithms," in *2022 4th Int. Conf. on Advances in Computing, Communication Control and Networking (ICAC3N)*, IEEE, 2022, pp. 242–248.

[5] M. C. Hung, P. H. Hsia, X. J. Kuang, and S. K. Lin, "Intelligent portfolio construction via news sentiment analysis," *International Review of Economics & Finance*, vol. 89, pp. 605–617, 2024.

[6] Y. Chen, "Financial news sentiment analysis method based on WMSA-Bi-LSTM," in *2023 4th International Conference on Intelligent Design (ICID)*, IEEE, 2023, pp. 319–322.

[7] A. Anderson and P. Brockman, "An examination of 13F filings," *Journal of Financial Research*, vol. 41, no. 3, pp. 295–324, 2018.

[8] T. Seth, et al., "A unified framework to assess market implications of institutional investments," in *2022 IEEE Int. Conf. on Big Data (Big Data)*, IEEE, 2022, pp. 1914–1921.

[9] L. Angelini, et al., "Systematic 13F hedge fund alpha," *SSRN*, 2019.