Causal Discovery in Stock Return

Xinqi Huang xih037@ucsd.edu
Ruizhe Dai rdai@ucsd.edu
Vivian Zhao vxzhao@ucsd.edu
Yishan Cai yic075@ucsd.edu
Biwei Huang bih007@ucsd.edu
Jelena Bradic jbradic@ucsd.edu

Abstract

Many people who own stocks don't have enough financial knowledge to understand the stock market. This project aims to empower investors by identifying the key drivers behind stock returns, providing actionable insights to guide decision-making and mitigate risks. To achieve this, we develop a multi-model framework that integrates sentiment analysis, historical stock data, and macroeconomic indicators. Public sentiment, a crucial factor in stock market dynamics, is analyzed using FinBERT on company-specific tweet data. Historical stock performance is modeled using DeepAR, while macro and microeconomic factors, such as GDP, CPI, and company reports, are processed through a nowcasting model. Feature selection is performed using causal learning techniques, and the outputs of these models are synthesized via a fusion layer to produce comprehensive predictions. By combining diverse data sources and advanced modeling techniques, this project aims to offer a clear and accessible understanding of the factors influencing stock returns, supporting better investment decisions.

Code: https://github.com/VivianZhao12/CAPSTONE-stockreturn

1 Introduction

1.1 Introductory Paragraph

Daily Stock return prediction is a crucial aspect of financial analysis, heavily influenced by a complex interplay of macroeconomic factors, microeconomic trends, and investor anticipation. Since the end of 2019, global policies, economic disruptions, and socio-political tensions have introduced significant volatility into the stock market, impacting millions of investors worldwide. On April 29, 2022, Amazon shares plummeted 14% following a 5% surge the previous day, marking their steepest decline since 2006. This instance highlights the challenges of understanding and predicting stock movements. To better analyze market volatility, we focus on the most impacted industries—Technology, Retail, and Health—by analyzing highly volatile stocks, including Amazon (AMZN), Alphabet (GOOG), AT&T (T), CVS (CVS), Amgen (AMGN), and Abbott Laboratories (ABT).

Traditional forecasting models, such as factor models and autoregressive approaches, primarily rely on analyzing historical daily patterns using predefined economic relationships, which limits their ability to capture sudden shifts driven by external shocks and sentiment changes. Deep learning-based approaches, such as DeepAR, a neural network-based model, have emerged as powerful alternatives, offering the ability to bring in diverse factors for complex dependencies modeling, and are able to adapt to rapid shifts in trends. However, its black-box nature limits interpretability, making it difficult for investors and analysts to understand the reasoning behind its predictions. This lack of transparency often overshadows its predictive performance, reducing trust in model-driven financial decision-making.

In this paper, we propose an advanced stock return prediction framework that integrates PCMCI (Peter and Clark Momentary Conditional Independence), a causal discovery algorithm, with DeepAR to enhance both interpretability and accuracy. Our approach leverages PCMCI for causal structure discovery in our multivariate time-series data, identifying true causal relationships rather than relying on correlations. The approach involve two steps: 1) Feature Selection based on causal relevance to filters out spurious correlations and retains the most influential factors, ensuring the simplicity of model for integrability 2) Lag Optimization to determine the optimal time dependency, providing clear decision rationales in our model for transparency.

Additionally, we incorporate external factors at varying granularities using a nowcasting model—a real-time economic forecasting approach widely used by central banks—alongside financial sentiment analytics, with a learned weighting layer for our final forecast. By integrating causal inference techniques with deep learning and real-time forecasting models, our framework bridges the gap between predictive power, interpretability, and data availability lag, making advanced stock forecasting more reliable and transparent for investors.

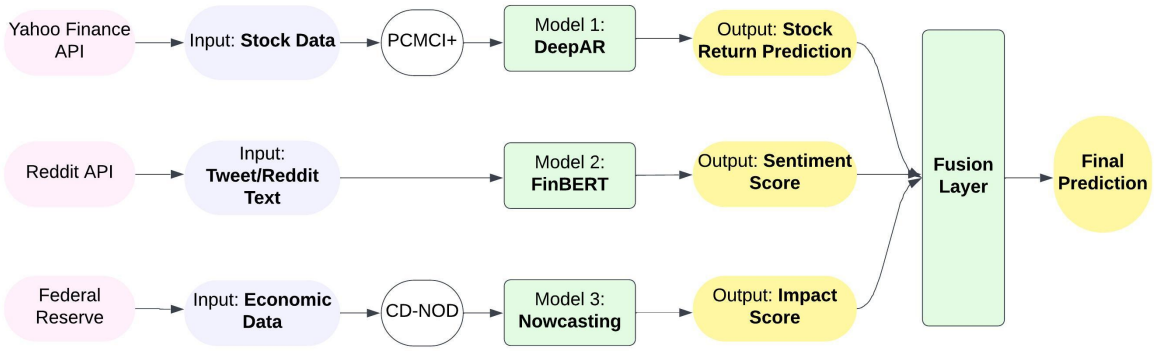


Figure 1: Overview of the Proposed Stock Return Prediction Framework

1.2 Literature Review

The challenge of identifying causal relationships in stock market returns presents unique complexities that extend beyond traditional statistical analysis. While numerous studies have examined correlations between market factors and stock performance, establishing genuine causal relationships remains elusive due to the dynamic, interconnected nature of financial markets and the presence of numerous confounding variables.

Prior research has established various approaches to understanding stock market predictability, focusing on both macroeconomic variables and media sentiment analysis. On the macroeconomic front, Engle, Ghysels and Sohn (2013) highlighted that fundamental factors such as industrial production growth, interest rates, inflation, and unemployment are key determinants of stock market movements, with Rapach, Wohar and Rangvid (2005) demonstrating through a comprehensive study of 12 industrialized countries that interest rates serve as the most reliable predictor in an international context. Parallel to these macroeconomic studies, research has increasingly recognized the importance of media sentiment in predicting stock returns. Notably, Heston and Sinha (2017) analyzed nearly one million news articles and found that daily news coverage could predict stock returns within a one to two-day window, with positive news generating immediate price responses while negative news produced delayed effects. This dual influence of macroeconomic factors and media sentiment was further reinforced by studies such as GARCÍA (2013) and Chen et al. (2014), who found significant relationships between media tone and future stock returns, particularly during periods of economic uncertainty.

There are also prior research works that have demonstrated the usage of DeepAR algorithm on predicting stock, Li et al. (2024); Xie, Lang and Liu (2023), they use the DeepAR algorithm to capture the pattern hidden in time series data, and their works have demonstrated the potency of the DeepAR algorithm in predicting under a complex environment, for example, the financial data. However they did not include the financial sentiment factor into their work. Moreover, there are also works that have demonstrated the effectiveness of using the nowcasting algorithm in solving the unmatched frequencies in macroeconomic data.

For example, Yiu and Chow (2010) used a nowcasting algorithm in predicting China's GDP and discovered that interest rate was extremely important in estimating China's GDP.

2 Methods

2.1 Stock Return Prediction Module

2.1.1 Data Collection and Preprocessing

We collected historical stock data for six major companies, with three from the technology sector and three from the healthcare sector using the Yahoo Finance API. The dataset spans from December 2019 to January 2025, providing 1,285 trading days per company, totaling 7,710 records.

For each company, we gathered daily metrics including opening price, closing price, adjusted closing price, daily high, daily low, and trading volume. These comprehensive metrics provide a complete picture of daily trading activities and price movements.

The raw data underwent several preprocessing steps. First, we structured the data chronologically and grouped it by company ticker to maintain consistent time series organization. To address the challenge of missing values that commonly occur in financial time series due to non-trading days or data collection issues, we implemented a two-stage imputation approach. The method begins with forward filling, which propagates the last valid observation forward to fill temporal gaps, preserving the time series characteristics. Any remaining missing values are then filled with zeros to ensure data completeness for model training.

For our prediction target, we computed daily returns using the percentage change in adjusted closing prices. The daily return at time t is calculated as:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

where R_t represents the daily return at time t, and P_t represents the adjusted closing price at time t. This metric serves as our target variable for the prediction task, capturing the daily price movements while accounting for stock splits and dividend payments through the use of adjusted closing prices.

The percentage changes of stock return in our datasets range from a -16% decline in CVS on May 1, 2024, to a 13% increase in Amazon on February 4, 2022.

For our analysis, we define changes below 4% as normal fluctuations, changes between 4% and 7% as significant fluctuations, and changes above 7% as abnormal fluctuations. Our model is designed to capture significant and abnormal fluctuations, while excluding normal fluctuations, in order to focus on more impactful price movements and avoid overfitting on minor patterns.

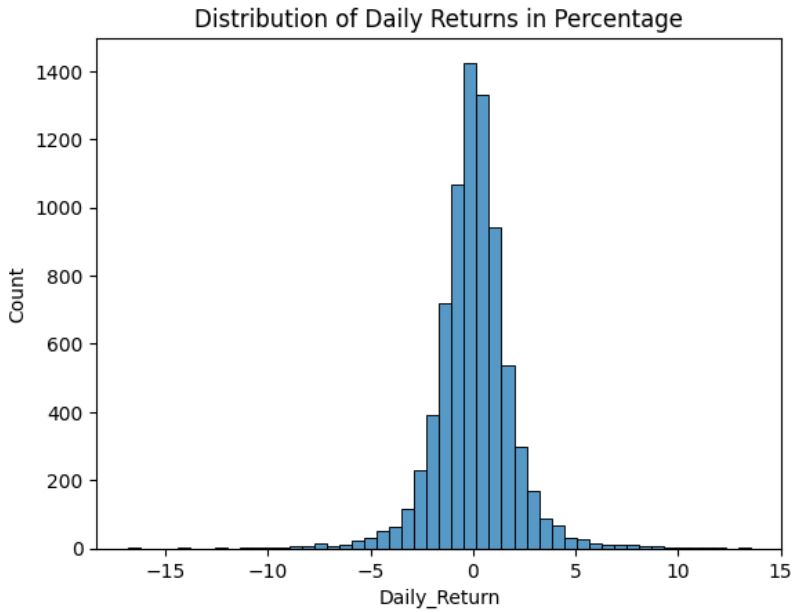


Figure 2: Distribution of Daily Returns in Percentage

2.1.2 Model Overview

For stock return prediction, we propose an enhanced DeepAR (Deep Auto-Regressive) architecture that combines deep learning with probabilistic forecasting. Our model is specifically designed to address three key challenges in stock return prediction: capturing long-term dependencies, handling multiple scales of price movements, and providing uncertainty estimates for risk management.

The model takes historical price data and carefully selected features as input, processes them through a deep neural network architecture, and outputs a probability distribution for future returns. This probabilistic approach not only provides point estimates but also quantifies the uncertainty of predictions, which is crucial for financial decision-making.

2.1.3 Feature Engineering and Selection

Our feature engineering process focused on creating a comprehensive set of predictive signals while avoiding redundancy and noise. We employed the PCMCI+ (Peter and Clark Momentary Conditional Independence with Plus) algorithm to perform causal feature selection, identifying the most relevant predictors while controlling for spurious correlations.

The selected features fall into three main categories:

1. **Technical Indicators**: Including moving averages, MACD (Moving Average Convergence Divergence), and volatility measures. These features capture price momentum and market dynamics.

2. **Market Microstructure Features**: Trading volume, bid-ask spreads, and intraday price

movements that reflect market liquidity and trading behavior.

3. **Temporal Features**: Calendar-based features including day-of-week and month-of-year effects, encoded using cyclic transformations to preserve their periodic nature.

2.1.4 Model Architecture

Our enhanced DeepAR architecture consists of four major components to produce accurate and reliable predictions:

Input Processing Layer The input layer of our model processes multivariate time series data through three main streams: daily historical returns, technical and market indicators, and temporal features. The input vector at each time step t is represented as $x_t \in \mathbb{R}^d$, with dimensionality $d = 1 + d_{cov} + d_{emb}$ comprising the previous day's return (z_{t-1}), a set of 8 covariate features (d_{cov}) selected through PCMCI+ causal discovery, and learned stock embeddings (d_{emb}) that capture entity-specific patterns. To ensure consistent scale across all features, we apply z-score standardization: $x_{normalized} = \frac{x-\mu}{\sigma}$.

Enhanced LSTM Core The core of our model utilizes an enhanced LSTM architecture with the following improvements:

- **Hierarchical Structure**: Multiple LSTM layers process the input at different temporal scales, allowing the model to capture both short-term fluctuations and long-term trends.

- **Skip Connections**: Residual connections between layers help maintain gradient flow and enable the model to bypass the deep network when simpler patterns are detected.

- **Orthogonal Weight Initialization**: For recurrent weight matrices, we employ orthogonal initialization, which helps maintain consistent gradient magnitudes during backpropagation through time. Specifically, for weight matrices connecting hidden states, orthogonal initialization ensures that the singular values of the weight matrices are all 1, effectively preventing gradient vanishing or explosion in the recurrent layers. This property is particularly crucial for financial time series, where capturing long-term dependencies is essential for accurate prediction.

- **Forget Gate Bias Initialization**: We implement a specialized initialization scheme for the LSTM forget gate bias, setting it to 1 while initializing all other biases to 0. This approach enables better gradient flow through time by allowing the network to learn long-term dependencies more effectively during the early stages of training. The forget gate's bias initialization is particularly crucial in financial time series analysis, where the model needs to learn to maintain or discard information over varying time horizons based on market conditions.

- **Variational Dropout**: We implement a temporal-consistent dropout mechanism that applies the same mask across time steps, preserving temporal coherence while preventing overfitting.

Attention Mechanism To improve the model's ability to focus on relevant historical patterns, we incorporate a temporal attention mechanism:

$$\alpha_t = \text{softmax}(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a))$$

where α_t represents attention weights for each time step. This allows the model to dynamically weight the importance of different historical time points when making predictions, particularly useful for capturing market regime changes and evolving relationships between features.

Probabilistic Output Layer Instead of producing single-point predictions, our model outputs a probability distribution for future returns. This is implemented as a Gaussian distribution whose parameters (mean and variance) are computed by the network:

$$p(z_t|z_{<t}, x_t) = \mathcal{N}(\mu_t, \sigma_t^2)$$

where $\mu_t = W_\mu h_t + b_\mu$, $\sigma_t = \text{softplus}(W_\sigma h_t + b_\sigma)$. This probabilistic approach provides not just predictions but also confidence intervals, enabling better risk assessment and decision-making.

2.1.5 Training Methodology

The model is trained using a carefully designed composite loss function that addresses multiple aspects of financial prediction:

$$\mathcal{L}total = \mathcal{L}likelihood + \lambda_1 \mathcal{L}small + \lambda_2 \mathcal{L}pattern$$

1. **Maximum Likelihood Estimation**: The primary component is the negative log-likelihood of the predicted distributions:

$$\mathcal{L}_{likelihood} = -\mathbb{E}[\log p(z_t|\mu_t, \sigma_t)]$$

encouraging accurate probabilistic forecasts.

2. **Small Movement Detection**: A specialized component of the loss function gives additional weight to capturing small price movements:

$$\mathcal{L}_{small} = \mathbb{E}[w_t|\Delta\mu_t - \Delta z_t|]$$

where $w_t = e^{-2|\Delta z_t|}$ weights smaller changes more heavily. This is because small movements are challenging to predict but crucial for trading strategies.

3. **Pattern Consistency**: We include a pattern matching term:

$$\mathcal{L}pattern = \text{MSE}(\text{normalize}(\mu t : t + w), \text{normalize}(z_{t:t+w}))$$

that ensures predictions maintain the statistical properties observed in historical data.

The training process employs the Adam optimizer and implements early stopping based on validation set performance (measured by Normalized Deviation) to prevent overfitting. The best model weights are saved during training when validation performance improves.

2.1.6 Model Evaluation

We evaluate our model using metrics specifically chosen for financial time series prediction:

1. **Normalized Deviation (ND)**: Measures prediction accuracy while accounting for the scale of price movements:

$$\text{ND} = \frac{\sum_t |\hat{r}_t - r_t|}{\sum_t |r_t|}$$

where $\hat{r}_t$ and r_t are predicted and actual returns.

2. **RMSE**: Captures the magnitude of prediction errors:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{r}_t - r_t)^2}$$

3. **Quantile Loss**: Evaluates the quality of the probabilistic forecasts:

$$\text{QL}\tau(y, \hat{y}) = \sum t(\tau(y_t - \hat{y}t) + +(1-\tau)(\hat{y}t - y_t)+)$$

where τ is the quantile level.

2.2 Sentiment Analysis Module

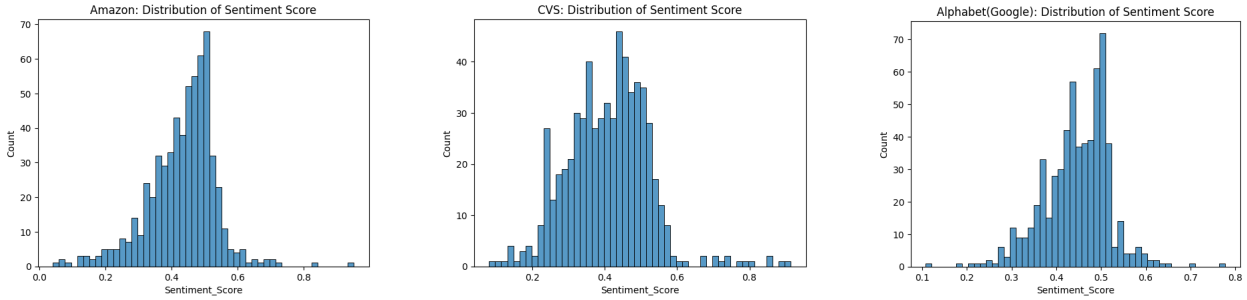
2.2.1 Data Collection and Preprocessing

We believe that incorporating sentiment factors could help our model to capture more information and have better performance on predicting stock returns. To make the numerical conversion of raw tweet data more accurate, we adopted FinBERT, a pre-trained NLP model to analyze sentiment of financial text. We have found a dataset on GitHub repository that contains tweets about famous companies including: Apple, CVS, Ebay, dating from 2020-6-01 to 2023-5-31. However, our goal is to make our model more applicable, which we need more recent tweets to make our model better capture the pattern in the recent stock market. Initially we thought about scraping more tweets, but the cost of accessing X API was too costly for our group. Instead, we decided to use Reddit posts and comments, as Reddit's API is free. One obstacle is that Reedit's posts and comments are sporadic, and there are cases that in some day, there were no posts or comments discussing stocks of companies we want to scrape. Therefore, after using FinBERT to convert raw text regarding each company's text to numerical sentiment scores, we decided to use a linear interpolation method that is scaled by time differences to alleviate the irregular spacing of our sentiment data scraped from Reddit and also making the interpolation of our missing data more realistic.

2.2.2 Score Computation

Using FinBERT's sentiment labels and their associated confidence levels, we computed our final sentiment score using the weighted sum of the sentiment labels (positive, negative,

and neutral), where each label was weighted by its corresponding confidence level. These sentiment scores were then normalized to a range between 0 and 1 with an approximately normal distribution.



(a) AMZN (b) CVS (c) GOOG

Figure 3: Sentiment Score Distribution of Stocks

2.3 Economic Impact Analysis Module

2.3.1 Data Extraction

For this project, we extracted microeconomic data using the Yahoo Finance API. The microeconomic data includes features such as the number of treasury shares, ordinary shares, net debt, sales of business. These features capture company-specific financial metrics crucial for understanding stock return dynamics.

On the macroeconomic side, data was sourced from the Federal Reserve of St. Louis (FRED), including indicators like Money Supply (M1 and M2), Interest Rates, Producer Price Index (PPI), Real Dollar Index, Unemployment Rate, Consumer Price Index (CPI), and Gross Domestic Product (GDP). Together, these datasets form a comprehensive view of economic and company-specific factors impacting stock movements.

2.3.2 Frequency Alignment

One of the main challenges encountered during the data preparation phase was aligning the data frequencies. Due to the nature of the datasets, GDP and company financial reports (microeconomic data) are only available on a quarterly basis, while most macroeconomic data is reported monthly. To address this, we implemented a frequency alignment process.

For each quarter, we identified the three monthly data points (e.g., January, February, and March for Q1) across all years and organized them into three separate vectors of time series data: one for the first month, one for the second month, and one for the third month of each quarter. We then performed regression analysis to predict quarterly stock returns using these vectors. The regression results provided weights for each monthly vector, which were then subjected to hypothesis testing to assess their statistical significance. This process

determined which month or combination of months best represented the quarterly trends. Based on these findings, the most representative data points were selected and aggregated to construct the quarterly datasets.

2.3.3 Feature Selection

With the aligned datasets, we performed feature selection using causal learning techniques. Specifically, we employed the CD-NOD algorithm to identify the most relevant predictors for stock returns while minimizing redundancy. This method allowed us to focus on features with significant causal relationships to the target variable, thereby enhancing the interpretability and performance of the nowcasting model.

2.3.4 Nowcasting Implementation

The final step in this segment of the pipeline involves applying the nowcasting model to predict stock returns using the selected features. Nowcasting, a real-time forecasting approach, leverages the latest available data to provide immediate predictions. By incorporating both macroeconomic indicators and company financial metrics, the model delivers a holistic view of stock return drivers. This integration allows us to capture the complex interplay between broader economic trends and individual company performance, providing more accurate and actionable insights.

Through this multi-step process, we ensure that the data feeding into the nowcasting model is both robust and relevant, enabling precise stock return predictions that account for a wide array of influencing factors.

2.4 Fusion Layer

2.4.1 Design Philosophy

2.4.2 Implementation

2.4.3 Weight Calibration

3 Results

3.1 Sentiment Analysis Module

3.2 Stock Return Prediction Module

3.3 Economic Impact Analysis Module

4 Discussion

5 Conclusion

6 Contributions

Xinqi mainly worked on collecting stock data□microeconomics data and sentiment data. She also did research and proposed method to help refine the pipeline of our model. In the report, Xinqi was responsible for the introduction paragraph, part of the data section of stock data and graphs.

Jason mainly worked on collecting sentiment data and interpolating for missing data. In the report, he was responsible for writing the literature review and sentiment analysis module.

Vivian contributed to verifying the usability of the DeepAR algorithm, collecting macroeconomic data, and working on the data frequency mismatch issues, transforming monthly macroeconomic data into quarterly datasets using a dynamic factor model.

Yishan was responsible for model design of the FinBERT sentiment analyzer, PCMCI feature selection process, as well as the implementation and optimization of the DeepAR prediction model. In the report, she was responsible for the method of stock return prediction module.

7 Acknowledgments

References

Chen, Hailiang, Prabuddha De, Yu (Jeffrey) Hu, and Byoung-Hyoun Hwang. 2014. "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media." [Link]

Engle, Robert F., Eric Ghysels, and Bumjean Sohn. 2013. "STOCK MARKET VOLATILITY AND MACROECONOMIC FUNDAMENTALS." *The Review of Economics and Statistics* 95 (3): 776–797. [Link]

GARCÍA, DIEGO. 2013. "Sentiment during Recessions." [Link]

Heston, Steven L., and Nitish Ranjan Sinha. 2017. "News vs. Sentiment: Predicting Stock Returns from News Stories." Routledge. [Link]

Li, Jiacheng, Wei Chen, Zhiheng Zhou, Junmei Yang, and Delu Zeng. 2024. "DeepAR-Attention probabilistic prediction for stock price series." Springer Science and Business Media LLC

Rapach, David E., Mark E. Wohar, and Jesper Rangvid. 2005. "Macro variables and international stock return predictability." [Link]

Xie, QingLin, Qi Lang, and Xiaodong Liu. 2023. "Stock Price Forecasting Based on Feature Fusion Deepar Model." In *2023 35th Chinese Control and Decision Conference (CCDC)*. [Link]

Yiu, Matthew S., and Kenneth K. Chow. 2010. "Nowcasting Chinese GDP: information content of economic and financial data." Routledge. [Link]

Appendices

A.1 Project Proposal

Causal Discovery in Stock Return

Xinqi Huang xih037@ucsd.edu
Ruizhe Dai rdai@ucsd.edu
Vivian Zhao vxzhao@ucsd.edu
Yishan Cai yic075@ucsd.edu
Biwei Huang bih007@ucsd.edu
Jelena Bradic jbradic@ucsd.edu

1 Introduction

1.1 General Theme

Many people who own stocks don't have enough financial knowledge to understand the stock market. This lack of knowledge makes it harder for them to avoid losing money, which could be important for their families. The stock market changes quickly and can be hard to predict, making it even more difficult to protect investments. We will spend 10 weeks to address this issue using causal inference algorithms on a dataset that tracks different factors and stock returns. By finding the main causes behind stock returns, we hope to give people a simple and clear explanation of what matters most in the stock market. This can help investors make better decisions and reduce the chance of losing money.

1.2 Problem Statement

The challenge of identifying causal relationships in stock market returns presents unique complexities that extend beyond traditional statistical analysis. While numerous studies have examined correlations between market factors and stock performance, establishing genuine causal relationships remains elusive due to the dynamic, interconnected nature of financial markets and the presence of numerous confounding variables.

Prior research has established various approaches to understanding stock market predictability, focusing on both macroeconomic variables and media sentiment analysis. On the macroeconomic front, Engle, Ghysels and Sohn (2013) highlighted that fundamental factors such as industrial production growth, interest rates, inflation, and unemployment are key determinants of stock market movements, with Rapach, Wohar and Rangvid (2005) demonstrating through a comprehensive study of 12 industrialized countries that interest rates serve as the most reliable predictor in an international context. Parallel to these macroeconomic studies, research has increasingly recognized the importance of media sentiment in predicting stock returns. Notably, Heston and Sinha (2017) analyzed nearly one million news articles and found that daily news coverage could predict stock returns within

a one to two-day window, with positive news generating immediate price responses while negative news produced delayed effects. This dual influence of macroeconomic factors and media sentiment was further reinforced by studies such as GARCÍA (2013) and Chen et al. (2014), who found significant relationships between media tone and future stock returns, particularly during periods of economic uncertainty.

Our Quarter 1 project demonstrated the effectiveness of causal discovery algorithms (PC, FCI, and GES) on a simulated dataset. However, applying these methods to real-world financial data presents additional challenges. Unlike simulated data, real financial markets exhibit intricate relationships between macro indicators (CPI, GDP, unemployment rate, policy interest rates), micro conditions (company financials, liability, cash flow), and market sentiment derived from social media data. These relationships often violate the assumptions of basic causal discovery algorithms. Furthermore, stock market data inherently includes temporal dependencies that weren't present in our Quarter 1 simulated dataset, necessitating the exploration of time-series-specific causal discovery methods.

To address these challenges, our research extends beyond the basic PC, FCI, and GES algorithms used in Quarter 1. We investigate more sophisticated approaches, including time-series extensions of FCI (tsFCI), neural network-based causal discovery methods, and hybrid approaches that combine traditional and modern causal inference techniques. This expansion of methodological tools allows us to better capture the complexity of real-world financial data. For the domain expert, our specific research question focuses on extending and adapting causal discovery algorithms to accurately identify the causal relationships between multiple market factors and stock returns. This investigation must account for temporal dependencies, hidden confounders, non-linear relationships, and the potential violation of causal sufficiency assumptions in financial market data. This approach addresses a critical gap in both causal inference methodology and financial market analysis, as previous work has either focused on simplified causal discovery in controlled settings or relied on traditional statistical analysis of market factors.

1.3 Output Expectation

In the next phase of our project, we aim to present our findings through a detailed and visually engaging poster. The poster will summarize the results of our causal discovery analysis, highlighting the primary drivers behind stock returns with DAGs. By combining textual explanations, visualizations of causal networks, and key data insights, the poster will serve as an accessible medium to convey our methodology and conclusions. We will focus on demonstrating the relationships between macroeconomic factors, company-specific metrics, and market sentiment, and the implications for financial decision-making. We will be validating these results with an Economics Department Professor and prior domain knowledge.

We choose the poster format for the following reasons:

1. **Visualizes Complexity**: Causal relationships are intricate, and diagrams or network graphs can convey these insights more effectively than textual descriptions.

2. **Engages Audience**: The format encourages viewers to explore sections of interest, enabling focused discussions and interactive learning.
3. **Supports Accessibility**: Summarized findings are easy to comprehend, making the content suitable for both technical and non-technical audiences.
4. **Fosters Feedback**: The conversational nature of poster sessions allows us to gather constructive criticism and perspectives, enriching our understanding and future work.

1.4 Potential Data

1.4.1 Variables Selection

To discover the causal relationship in stock return, we selected a set of variables that represent macroeconomic conditions, company-specific financial metrics, and market sentiment, as we hypothesized that these factors collectively influence stock prices.

- **Macro Conditions**:
  - Consumer Price Index (CPI): Reflects inflation trends, which are related to individual investor confidence.
  - Gross Domestic Product (GDP): Measures overall economic health and market conditions.
  - Unemployment Rate: Indicates labor market conditions and have direct affect on individual spending.
  - Interest Rate: Directly influences investment decisions.
  - Annual Growth in Interest Rate: Captures the trend of investment activity.
  - Geopolitical Events: Represents global stability, which can impact investor behavior.

  The macroeconomic condition reflects the overall health of the economy, market conditions, job market stability, and international relations which directly impact individual and institutional confidence in investing. These factors shape investor sentiment and risk tolerance, which in turn influence stock prices and returns.

- **Micro Conditions**:
  - Industry: The sector or market in which a company operates.
  - Dividend Rate: The proportion of earnings paid to shareholders as dividends.
  - Overall Risk: A composite measure of various risks.
  - Return on Equity (ROE): A measure of profitability, indicating how effectively a company uses shareholders' equity to generate profits.
  - Liabilities: The company's debts or financial obligations.
  - Cash Flow: The net amount of cash generated or used by a company.
  - Revenue Growth: The rate at which a company's sales are increasing.
  - Debt-to-Equity Ratio: A financial leverage ratio.

  The microeconomic conditions offer a detailed perspective on a company's financial health, its efficiency in generating profits for shareholders, and trends within the

industry that would influence investment decisions. These factors are closely linked to stock pricing and, consequently, have a significant relationship with stock returns.

- **Market Sentiment**:
  - Tweet Sentiment: Reflects the collective perception and emotional reaction of investors, influencing short-term stock price movements.

  In economic theory, people's anticipation of future events significantly influences their investment behavior. If investors believe that a company will perform well in the future, either based on their insights on earnings forecasts, new product launches by the company, or positive news reports, they are likely to buy the stock and thus drive its price up. Moreover, when a large number of investors act on sentiment, their collective actions can create a feedback loop due to herd behavior. As a result, we see the necessity of including sentiment as one of our factors. We would include the tweet sentiment on companies to find the causal relationship to the stock return as tweets provide real-time, unfiltered insights into market sentiment, reflecting the emotional and speculative reactions of a diverse investor base.

1.4.2 Data Sources

- **Macroeconomic Variables**: These data can be easily scraped or collected from reliable institutions such as the *U.S. Bureau of Labor Statistics*, *Bureau of Economic Analysis*, and *Federal Reserve Board*. Given that these data are typically reported on a monthly, quarterly, or annual basis, they can also be obtained manually if needed.
- **Microeconomic Variables**: These data can be obtained using the *Yahoo Finance API*, which provides comprehensive company information, including equity, liabilities, and dividends paid to stockholders. The source is reliable as Yahoo Finance's API is one of the pioneers in the financial data market.
- **Market Sentiment and Stock returns**: We will conduct sentiment analysis on tweets from the following dataset: Tweet Sentiment's Impact on Stock Returns.

  This dataset contains 862,231 labeled tweets along with associated stock returns, offering a comprehensive view of how social media sentiment affects company-level stock market performance. The dataset includes the following variables:
  - **Tweet Text**: The content of the tweet.
  - **Stock Symbol**: The company ticker associated with the tweet.
  - **Date**: The date the tweet was posted.
  - **Closing Price at the Time of Tweet**: The stock's closing price when the tweet was published.
  - **10-Day Volatility**: Stock price volatility over the subsequent 10 days.
  - **30-Day Volatility**: Stock price volatility over the subsequent 30 days.

  The tweet data have been directly extracted from the platform and are ready for analysis after pre-processing. Also, the volume of data available in the dataset helps mitigate the effects of missing data and potential information loss during sentiment analysis. For any missing stock return data, we will supplement it using the Yahoo

Finance API to ensure completeness.

References

Chen, Hailiang, Prabuddha De, Yu (Jeffrey) Hu, and Byoung-Hyoun Hwang. 2014. "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media." [Link]

Engle, Robert F., Eric Ghysels, and Bumjean Sohn. 2013. "STOCK MARKET VOLATILITY AND MACROECONOMIC FUNDAMENTALS." *The Review of Economics and Statistics* 95 (3): 776–797. [Link]

GARCÍA, DIEGO. 2013. "Sentiment during Recessions." [Link]

Heston, Steven L., and Nitish Ranjan Sinha. 2017. "News vs. Sentiment: Predicting Stock Returns from News Stories." Routledge. [Link]

Rapach, David E., Mark E. Wohar, and Jesper Rangvid. 2005. "Macro variables and international stock return predictability." [Link]