# Related work (Peter, Vijay, Andrew, Harrison)

Copy from this report.

## Abstract:

This report investigates correlations between companies to identify potential partnerships, predict future stock movement, and discuss across multiple industries. By leveraging public SEC filings and stock prices via yFinance API. Our analysis uses correlation, clustering techniques and integrates sentiment analysis to LSTM models. This multidimensional approach reveals corporate dynamics, evaluates business decisions, and uncovers significant insights.

## Introduction:

### Problem statement:

Understanding intercompany dynamics in the stock market can give valuable insights into partnerships and performance trends. However, current approaches fail to establish a clear connection between corporate actions and stock performance, leaving investors and stakeholders with fragments of information.

### Why the problem matters:

In the corporate world, where companies default every day and business landscape shift rapidly, understanding intercompany dynamics can feel like navigating a labyrinth. The stock market is a complex web of partnerships and performance trends, where corporate actions ripple through the financial ecosystem. However, existing methods often fall short in comprehensively linking these corporate actions to stock movements, leaving critical untapped insights. Business decisions wield immense influence over stock performance, shaping outcomes for investors and stakeholders. In this report we want to bridge the gap between textual corporate data, such as 10-K reports and numerical stock data.

### Challenges:

The road to achieving this fraught with difficulties includes:

* Extracting meaningful insights from vast and unstructured textual datasets.
* Ensuring interpretability in clustering methods to reveal coherent patterns.
* ..

### Key Ideas:

Our approach leverages advanced techniques such as sentiment analysis, keyword clustering, and decision trees to link corporate language to stock performance. By doing so, we aim to enable predictive analytics and valuable insights that drive the stock market.

## Related Work:

### Existing Research:

Current literature on stock movement prediction primarily focuses on numerical data like stock prices. Limited work exists on integrating textual data, such as 10-K reports, for predictive insights.

### Limitations:

Previous methods struggle with processing unstructured textual data, especially at the scale of SEC filings. Additionally, most clustering models lack clarity in interpreting results for corporate strategies.

### Our Contribution:

By combining sentiment analysis, correlation studies, and clustering, our work overcomes date extraction challenges and provides actionable insights into corporate dynamics.

## Methodology

### Datasets, tools, main tasks, analytical thinking (Peter, Vijay, Andrew, Harrison)

### Datasets and Tools:

* Datasets:
  + SEC 10-K, 10-Q, and 8-K reports (~~25 GB) sourced from [EDGAR](https://www.sec.gov/edgar/search/) .
  + Stock prices through Yahoos yFinance API([yFinance](https://pypi.org/project/yfinance/)).
* Tools Used:
  + Python (libraries for downloaders, extractions, sentiment analysis, and modeling), t-SNE, K-means,..)

We gathered and processed around 26GB of data for edger sec Edgar Database: <https://www.sec.gov/edgar/search>

### Main Tasks:

1. Data gathering: Gather filings from reliable sources and create a data lake.
2. Clustering: Grouping companies by keywords and sentiment …
3. Prediction: Find meaning full correlation between companies

### Analytical Thinking

* Identified patterns connecting corporate keywords to stock trends.
* Expected strong industry specific correlation but found surprises like NVIDIA and intel clustering separately.

### Data Understanding and preprocessing:

* Performed exploratory analysis on SEC filings to identify key trends and frequent keywords.
* Created a relational sqlite3 database with key value pairs of data to distinguish individual companies with specific tags as below (abbreviated in final slides)
  + RF
  + URFC
  + Cybersecurity
  + Properties
  + MDAFC
  + QQDMR
  + FSSD

### Discussion

What worked well

As we progressed in developing our tools, we observed a strong correlation between NVIDIA and ADP. Intrigued by this finding, we delved deeper to understand the reasons behind their connection. During our research, we came across an article that validated our results: [AMD Earnings Show Power of NVIDIA Software, ADP Data Defies](https://www.moomoo.com/news/post/45207411/amd-earnings-show-power-of-nvidia-software-adp-data-defies?level=2&data_ticket=a1f55cb7ca194eb87fb62f96c938d1d6).

#### Challenges

* Extracting information from 10-K reports posed significant challenges due to their inconsistent formatting and unstructured nature
* Processing quarterly reports proved difficult because of computational constraints, as the large volume of data required substantial processing power.
* To handle the massive datasets efficiently, we implemented multithreading, which significantly improved the downloading and processing speed.

# Appendix (Peter, Vijay, Andrew, Harrison)

We include the honor code pledge in the appendix to certify that we have neither violated nor concealed any violations of the University of Colorado – Boulder honor code.

``On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work. ‘’

Work done by each team member:

Peter: data understanding, clustering, sentiment analysis.

Vijay: Dataset collection, data preprocessing, database design.

Andrew: Evaluation, clustering analysis, model interpretation.

Harrison: Methodology, decision tree implementation, discussion.