# Methodology (Peter, Vijay, Andrew, Harrison)

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

For each tool used (Peter, Vijay, Andrew, Harrison)

* What are we using it for?
* Are there alternatives?
* If there are alternatives, why did we pick the tool we are using?

For sentiment analysis we used the Distilbert uncased model, which is machine learning based. It is part embedding and then a sentiment layer upon that. There are a lot of different options for sentiment analysis, but ultimately this is the one we settled on for a few reasons. The foremost being that we felt it give the best sentiment, taking into account the context of words. Another reason we used this model was the fact that it was fine-tuned on sentiments of proper English (i.e. no slang and such) which is what company documents like the 10-K are using.

For each main task (Peter, Andrew, Harrison)

* What is the main task?
* What was our initial approach to solving the main task?
* What difficulties did we find in solving the main task? How did these difficulties change our approach to solving the problem?
* What are the technical details of the method that we finally used?

Our main task here is to compute the relevant sentiments from our Distilbert model and use them for analysis and modeling. Our initial approach was to simply enter the entire document into the model, but due to context length issues that didn’t work. The second approach we tried and the one we settled upon was simply splitting the document by context length with a sliding window to preserve context from the previous split. There was likely some textual context lost here, but I think our window of around 25 tokens allowed us to still get accurate sentiment readings. The only other concern with the Distilbert sentiment analysis was that it required a significant amount of compute, but when we moved the context windows into batches it was able to complete all our documents in around 5 minutes.

Analytical thinking (Peter, Andrew, Harrison)

* What kinds of patterns did we initially expect to find?
* Why did we collect the data for specific stocks?
* How can we interpret the solution to each main task? (i.e. what is the practical implication of the solution)
* Are there any caveats or assumptions we need to make clear when giving the interpretations?

We expected to find that the overall document sentiments of poorly performing companies were mostly negative and that the companies performing well would have positive sentiments. We found this to be somewhat the case with most companies embellishing the reports. The overall negative sentiments did shine through though. We chose stocks that we thought represented the idea of positive poor and good performance. When interpreting our sentiment scores for our reports we get a value from 0-1 which we then scale to a distance of –1,1. This distance represents how close the document is to a negative or positive sentiment. It is important to note that it does not represent the strength of the sentiment.

## Data understanding, preprocessing, warehouse, modeling (Peter, Vijay, Andrew, Harrison)

Data understanding (Peter, Andrew, Harrison)

* What exploratory analysis did we do on the data (what tables/graphs did we make that don’t directly answer the main questions, but still give interesting information)?
* Why did we do the exploratory analysis that was listed?

N/A

Preprocessing (Peter, Andrew, Harrison)

* Did we need to clean any data (replace missing values, substitute values, etc.)?
* Why did we make the choices we made when cleaning data?
* Did we need to perform any normalization or other transformation of the data?

N/A except distance transform described above

Modeling (Peter, Harrison)

* What models did we make to answer the main questions?
* What inputs do the models take?
* What outputs do the models give?
* Why should the models give reasonable answers to the questions we are interested in?
* Did we make any assumptions about the form of the data which relate to the exact model we chose to use?
* What parameters of the model could affect the output of the model?

To model stock prices, we used an LSTM which took in historical stock data in 60-day windows and the corresponding sentiment of the previous year's 10-K as features. The model then outputs the next day's prediction of the stock. In our literature review we identified LSTMs as a reasonable approach to stock modeling. This, with a dense layer for embedding the sentiment within the prediction, seemed the best approach to combining the features for an accurate output. This of course assumes that the sentiment of the reports is related to the stock price which we think is more than reasonable. To train our model we use the Adam optimizer with a learning of 10^-3, over 5 epochs. For exact architecture please reference model.ipynb.

# Evaluation (Peter, Andrew, Harrison)

Metrics

* What are we measuring about the model – how do we know our model is good?
* What statistics can we state about the model quality?

To evaluate the model, we used MSE from the true stock prices as well as graphs. In the single stock case, the model performed well matching the true prices and overall movement of the stock.

* What can we learn generally from the results we got from the model?
* Are there any specific, interesting, or surprising findings from our results?

Overall, the model performed well in the single stock case, but training became unstable in the multi-stock case. As we can see in multistock.png and singlestock1.png. It is interesting that the model was unable to generalize to more than one stock. This suggests that we either need more training data or a larger model for multi-stock prediction. Another hypothesis is also that a general model for stock prediction is not feasible with this data.

# Discussions (Peter, Vijay, Andrew, Harrison)

* Are there any techniques that you learned about through the project that you didn’t know about?
* Are there any significant data mining principles that you got to experience in the project firsthand?

What worked well/what didn’t

* Is there any technique or part of the project which worked particularly nicely in the project?
* Why do you think that it worked well? How could you use this knowledge in the future?
* Is there any technique or part of the project which worked particularly badly?
* Why do you think that it worked poorly? How could you avoid or improve upon the deficiencies?

Directions for future work

* What would you want to do if you had another month on the project?
* What ideas do you have for exploring the data which are not directly related to the techniques we used in this project?

The LSTM model with sentiment analysis provided a lot of challenges for this project and for future work expanding the data it uses and refining the training method could make a great project on its own. The multi stock case is a new challenge and a direction we think future research could take as it performed particularly poorly. We do believe though that including the 10-K sentiment embedding within the model provides valuable information. Overall this project provided a great overview of the data mining process.