### **Slide 1: Introduction to the Problem**

The problem we're addressing is the inherent unpredictability of the financial markets, and the difficulty in forecasting stock market trends accurately. Accurate predictions can have significant impacts on investment decisions, portfolio management, and risk management strategies. So, the goal of our project was to create a model that could accurately predict future movements of the S&P 500 index. This could be used to trade ETFs such as SPY that are based on the index. The differentiating factor here is the inclusion of the component stock OHLC data for the S&P 500 index, resulting in over 3000 features per target observation.

### **Slide ?: Research and Approach**

In terms of methodology, we explored a machine learning-based approach to time-series forecasting. More specifically, we leveraged two different models:

1. **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with Exogenous Regressors)** - a time-series model that can handle seasonality and incorporates external features.
2. **Long Short-Term Memory (LSTM) Neural Networks** - a type of recurrent neural network (RNN) well-suited for sequential data, such as stock prices.

Both models aimed to predict the S&P 500 index using historical stock data, with a focus on top-performing stocks based on market capitalization. We also integrated technical indicators like stock opens, closes, and volume to enhance the predictive power of the models.

### **Slide 2: Findings from Research - Aman and Priyam**

During the research phase, we learned a few key things about the financial market predictions:

1. **Data Complexity**: Stock market data is highly volatile and affected by numerous unpredictable factors like economic news, geopolitical events, market sentiment as well as global pandemic like COVID. As a result, even state-of-the-art models often struggle to generate consistently accurate predictions.
2. **LSTM Model Performance**: The LSTM model performed reasonably well on short-term predictions, but the mean absolute error (MAE) was still high, indicating that predictions were not as accurate as desired. Despite this, the model tracked the index pretty well.
3. **SARIMAX Model Limitations**: The SARIMAX model, while traditionally strong for time series forecasting, was less effective here. The high dimensionality of the dataset meant that we had to limit the SARIMAX to train on the top 50 companies by market capitalization, this further degraded its performance.
4. **Model Comparison:**

| **Model** | **Mean Absolute Error** |
| --- | --- |
| LSTM | 105.1 |
| SARIMAX | 388.5 |

### **Slide 3: Implications of Results - Harsh and Izzy**

We ran a simple trading simulation using the predictions of both of our models. Starting with a thousand dollars and trading based on if the model predicts a increase or decrease in the index for the subsequent day. This simulation was run on data of about 500 days.

**Results:**

| **Model** | **Final Balance** | **% Profit** |
| --- | --- | --- |
| Naive | $1194.3 | 19.4 |
| LSTM | $1207.4 | 20.7 |
| SARIMAX | $1008.5 | 0.85 |

1. The naive result is the profit we would have made if we had simply bought and held the ETF for the entire duration of testing period, i.e it is the performance of the index itself
2. SARIMAX had insignificant profits in the simulation and overall performed quite poorly. The performance might have been improved with more training time.
3. LSTM marginally outperformed the market by about a single percentage point. However here we need to understand that the simulation did not consider the transactional costs that are associated with trades as well as assumed perfect liquidity and ability to perform trades at desired price. Factoring all of this in would certainly make the LSTM perform worse than the market itself.

**Future Improvements:**

1. **External Factors:** Our models currently do not consider the effects of external factors such as sentiment analysis, macroeconomic factors, or news events, which very much have an effect on the stock market. Having the factored into the model would improve its prediction performance.
2. **Model Complexity:** Increasing the model complexity for both LSTM and SARIMAX along with accommodating higher training times for hyperparameter tuning would also improve the prediction performance of the models.
3. **Technical Indicators**: Use advanced indicators like Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), or Bollinger Bands for better representation of market trends.

Overall, while we don’t beat the market outright, the approach of using component stock data to forecast the index values shows promise and can be further refined upon to create a successful trading strategy.