# BIG DATA

Luu Hoang Long Vo

Phu Hien Le

# Table of Content

**Part 1: Data Exploration**

**Part 2: Insights**

**Part 3:**

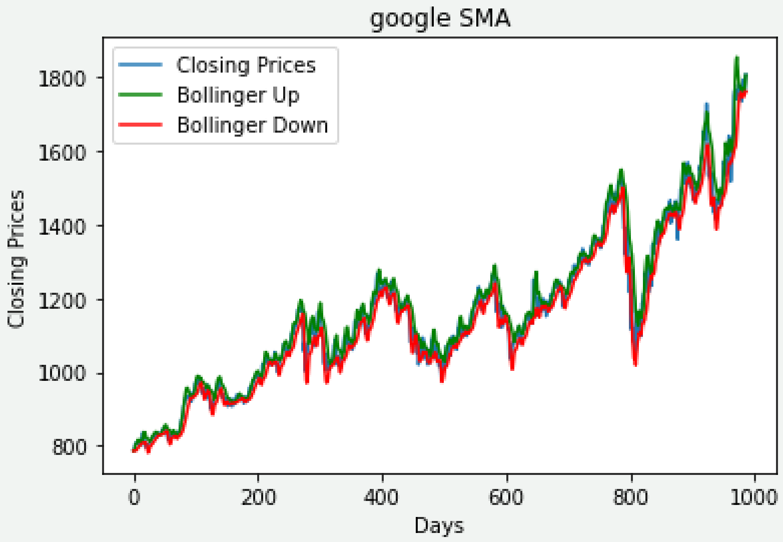
Machine Learning Pipeline
Results
Intepretation

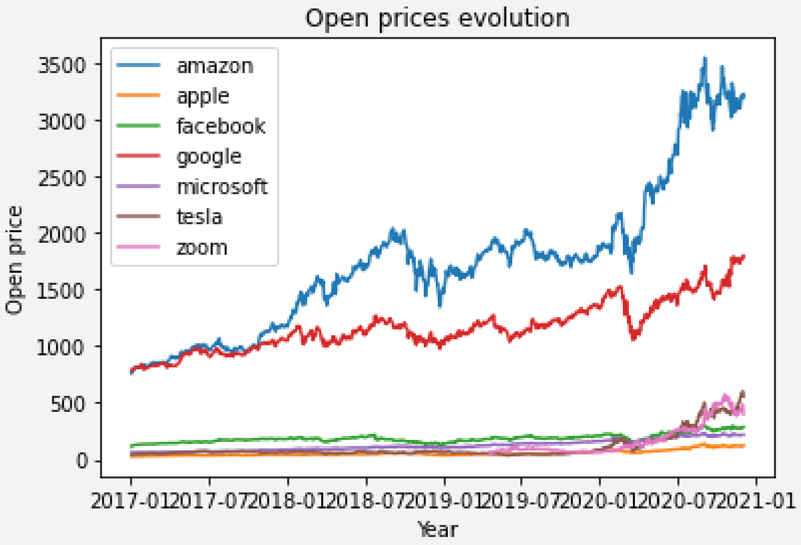
# Part 1: Data exploration

- Explore the Dataset with the given instructions:
  - Characterization of dataset
  - Missing values
  - Calculating key indicators
    - Highest daily return
    - Average for all stock price categories (daily, monthly, yearly)

# Part 2: Insights

- Evolution of Stocks
- Bollinger Bands





# Part 3: Machine Learning Pipeline

- Data cleansing
  - Imputer with strategies (replace with 0, drop the row, replace with mean)
- Data preparation
  - One hot encoded categorical variables (which in this case there's none)
  - Assembled the input as vector of double
- Estimator Selection (Linear Regression, Decision Tree, Random Forest Regression)
- Hyperparameter Tuning

# Part 3: Results

| | High | Low | Close | Open | AdjClose | Volume |
|---|---|---|---|---|---|---|
| RMSE | 26.95 | 28.26 | 34.45 | 24.24 | 34.35 | 3454448.40 |
| R2 | 0.98 | 0.97 | 0.96 | 0.98 | 0.96 | -2.31 |

Linear Regression

| | High | Low | Close | Open | AdjClose | Volume |
|---|---|---|---|---|---|---|
| RMSE | 107.4557 | 104.0034 | 107.904 | 105.1103 | 107.904 | 689368.414 |
| R2 | 0.6096 | 0.6521 | 0.612 | 0.6334 | 0.612 | 0.2195 |

Linear Regression (Log Scale)

# Part 3: Results

| | High | Low | Close | Open | AdjClose | Volume |
|---|---|---|---|---|---|---|
| RMSE | 112.4271 | 104.6856 | 109.8498 | 107.0157 | 107.0157 | 903302.823 |
| R2 | 0.5727 | 0.6476 | 0.5979 | 0.62 | 0.62 | -0.34 |

Decision Tree

| | High | Low | Close | Open | AdjClose | Volume |
|---|---|---|---|---|---|---|
| RMSE | 112.4729 | 104.7283 | 109.3667 | 107.0968 | 109.3667 | 873475.358 |
| R2 | 0.5723 | 0.6473 | 0.6014 | 0.6194 | 0.6014 | -0.253 |

Random Forest Regressor (Hyperparameter Tuning)

# Part 3: Intepretation

- All the stock price categories were able to be predicted with mediocre accuracy at best.
  - The future is inherently unpredictable
  - Does not account for qualitative variable (news, scandals, corruption)
- Volume was the hardest to predict.
  - Due to the unstable nature of buying and selling stocks
  - The focus of the market of a given day
- If it is easy to predict, everybody would be billionaire.