# Final project, option 1
# Time series prediction in Spark

- Data: use AAPL (Apple stock price information)
  - https://finance.yahoo.com/quote/AAPL/history?p=AAPL
- Download as .CSV file
- Use daily data for 1980 – 2018 (9478 records)
  - Feel free to leverage other data sources that can boost predictive power of your models. You can be creative in terms of what data you are using.
- Build a model for 'close' price prediction
- Bonus point for leveraging several features and building multivariate models
- Analyze prediction accuracy and report Mean Squared Error (RMSE) and sMAPE for predictions over varied horizons:
  - 1 day, 1 week, 2 weeks, 1 month, 4 months

Warning: Adding a portfolio optimization is not part of the final project. You can add it if its something you have time for, but its not a requirement, you will be judged by the quality of your models used for time series prediction and not portfolio optimization results.

# Final project
# Time series analysis

**Apply basic time-series forecasting along with more advanced methods:**

- Simpler methods:
  - Simple & Moving Average (SMA)
  - Exponential Smoothing
  - Autoregressive Integration Moving Average (ARIMA)

**More sophisticated ML methods:**

- Regressor Trees
- Ensembles of trees (XGB)

- You cant use Python based resources and libraries (they are not running on a cluster)

# Final project,
# Time series analysis

| Model | 1 day | 1 week | 2 weeks | 1 month | 4 months |
|---|---|---|---|---|---|
| Model 1 | sMAPE | sMAPE | sMAPE | sMAPE | sMAPE |
| Model 2 | sMAPE | sMAPE | sMAPE | sMAPE | sMAPE |
| Model 3 | sMAPE | sMAPE | sMAPE | sMAPE | sMAPE |

Test at least 2 models
Elaborate on the performance and what affects sMAPE for each model

Provide information about your models and how you built them:
Mention everything relevant to model construction and optimization process

| Model | Used data | Feature transformation | Train/test split | Overfittng? | Model Optimization |
|---|---|---|---|---|---|
| Model 1 | | | | | |
| Model 2 | | | | | |
| Model 3 | | | | | |

# Final project,
# Time series analysis

**Your report may be focused on the following questions:**

1) What dataset did you use and why did you choose it (if you chosen a dataset different from what has been offered)

2) What problem are you analyzing? (define business problem)

3) Comment on target variable

4) Literature review: what is being used for similar kind of analysis?

5) Explain why did you choose models and techniques you used? What is the rationale?

6) What metrics did you use to measure performance of the model? Why is this metrics relevant to the problem statement?

7) Present your results

7) How did you optimize your models?

8) Features: comment on feature importance and ranking. Any features that stood out that has strong influence on model’s predictive power?

6) What are challenges did you run into while building your model in Spark? Did you run into any limitations?

7) What were your mentor’s recommendations?

8) Overall recommendation to improve your results and this project

# Time series
# Support information

- Useful links:

https://databricks.com/session/time-series-analysis-with-spark

https://databricks.com/session/time-series-analytics-with-spark

ARIMA in Spark:

https://badrit.com/blog/2017/5/29/time-series-analysis-using-spark#.XI-W4VNKiL8