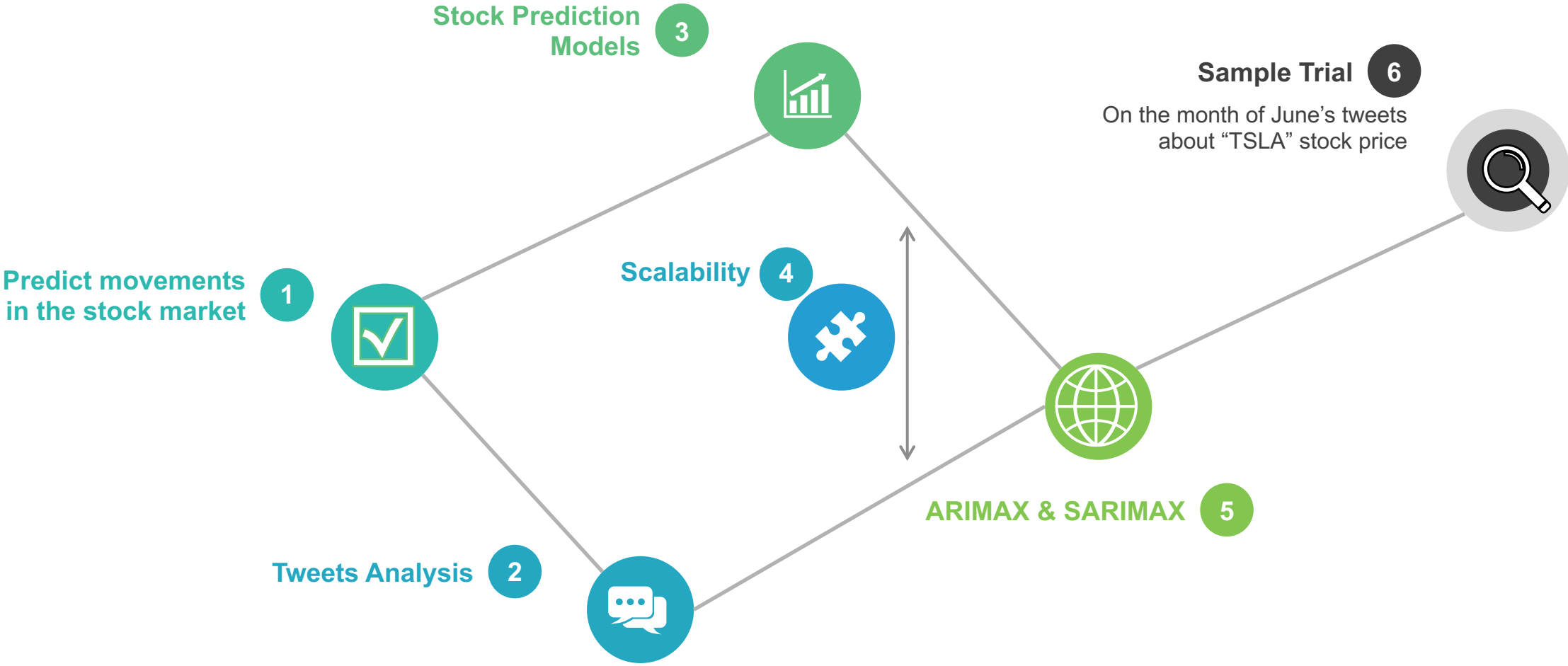
Stock Prediction

A stock prediction model approach with tweets sentiment analysis

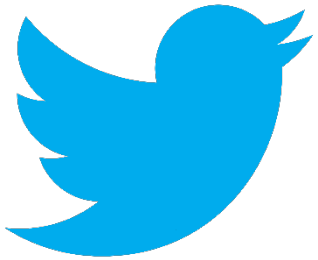
Giulio Mantoan, Giacomo Miolo, Mireia Riviere, Paolo Ticozzi

Professors Marco Brambilla and Danilo Ardagna
Scalable Data Storage and Processing
Business Analytics and Big Data

Overview


Stock Prediction Models 3
Sample Trial 6
On the month of June's tweets about "TSLA" stock price
Predict movements in the stock market 1
Scalability 4
ARIMAX & SARIMAX 5
Tweets Analysis 2

Data Sources

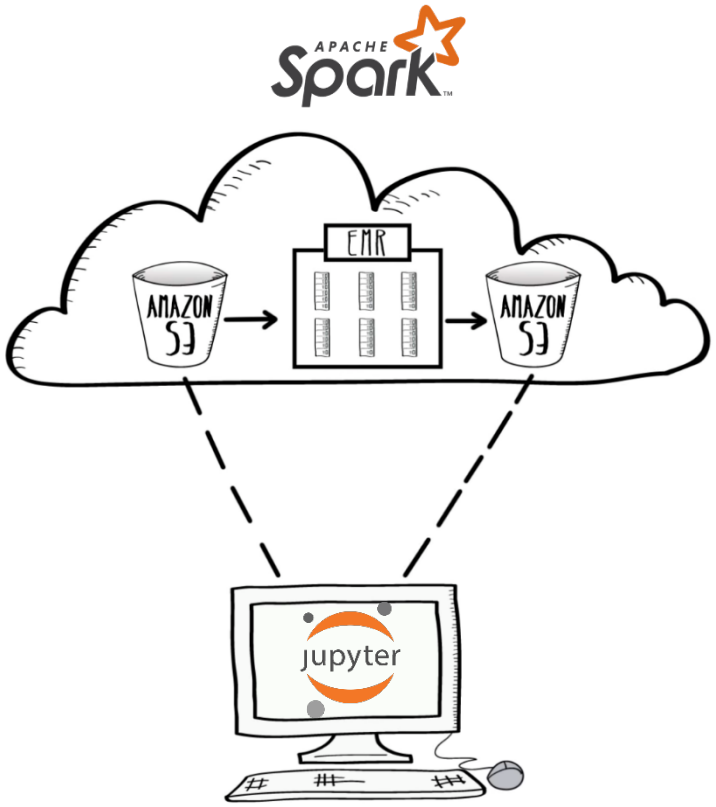


151 Million tweets
450 Gb

50 Experts Account

TSLA
historical stock price

Elastic Map Reduce



[4]:
```
df = spark.read.csv(
    "s3://aws-emr-resources-397910311797-eu-west-1/TSLA.csv", header=True, mode="DROPMALFORMED"
)
```

Spark Job Progress

Job [3]: csv at NativeMethodAccessorImpl.java:0

Progress for csv at NativeMethodAccessorImpl.java:0 — Job Progress: 1/1 Tasks Complete

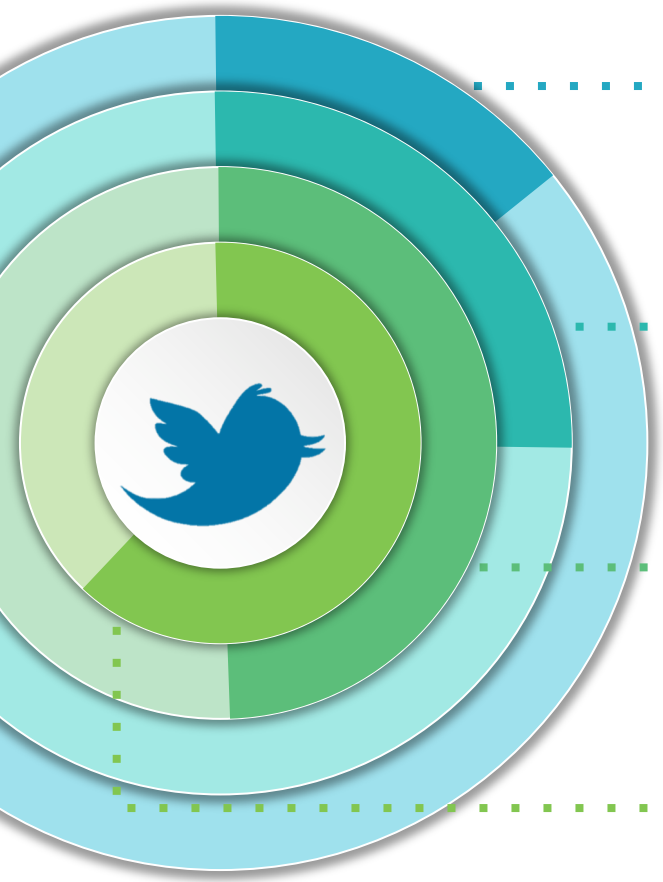
Stage [ID]: name at [source]:[lin...	Status	Task Progress	Elapsed Time (second...	Failed Task Logs
Stage [3]: csv at NativeMet...java:0	COMPLETE	1/1	1.273	

[22]:
```
df.show()
```

Spark Job Progress

```
+----------+----------+----------+----------+----------+----------+--------+
|      Date|      Open|      High|       Low|     Close| Adj Close|  Volume|
+----------+----------+----------+----------+----------+----------+--------+
|2015-02-09|215.380005|217.929993|211.990005|217.479996|217.479996| 3472400|
|2015-02-10|217.550003|220.500000|215.000000|216.289993|216.289993| 5390500|
|2015-02-11|212.210007|214.740005|207.279999|212.800003|212.800003| 9769100|
|2015-02-12|193.570007|203.089996|193.279999|202.880005|202.880005|15649600|
```

Tweet Structure



Dataset Fields

more than 150 fields for each tracked tweet, with records about user information, position, date and time, text, comments, retweets, ...

Fields Selection

we focused on:
UserID, User Name, Source, Geolocation and Coordinates, Place, Text, Lang, Date, Time

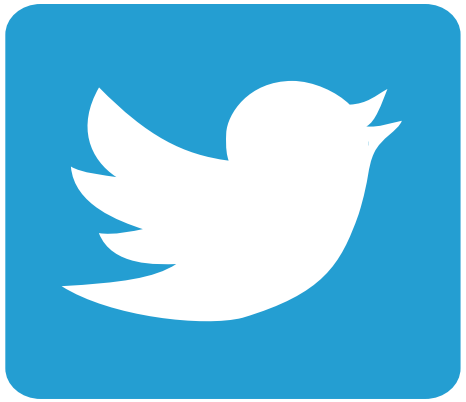
Tweets length

in Nov. 2017 the length of each tweet has been doubled from 140 characters to 280

DateTime

conversion from ISOdate to date and time. for simplicity, we considered just date, but the correct timestamp should have been in between 18.00 p.m of day 1 to 18.00 p.m. of day 2, due to stock market closing time

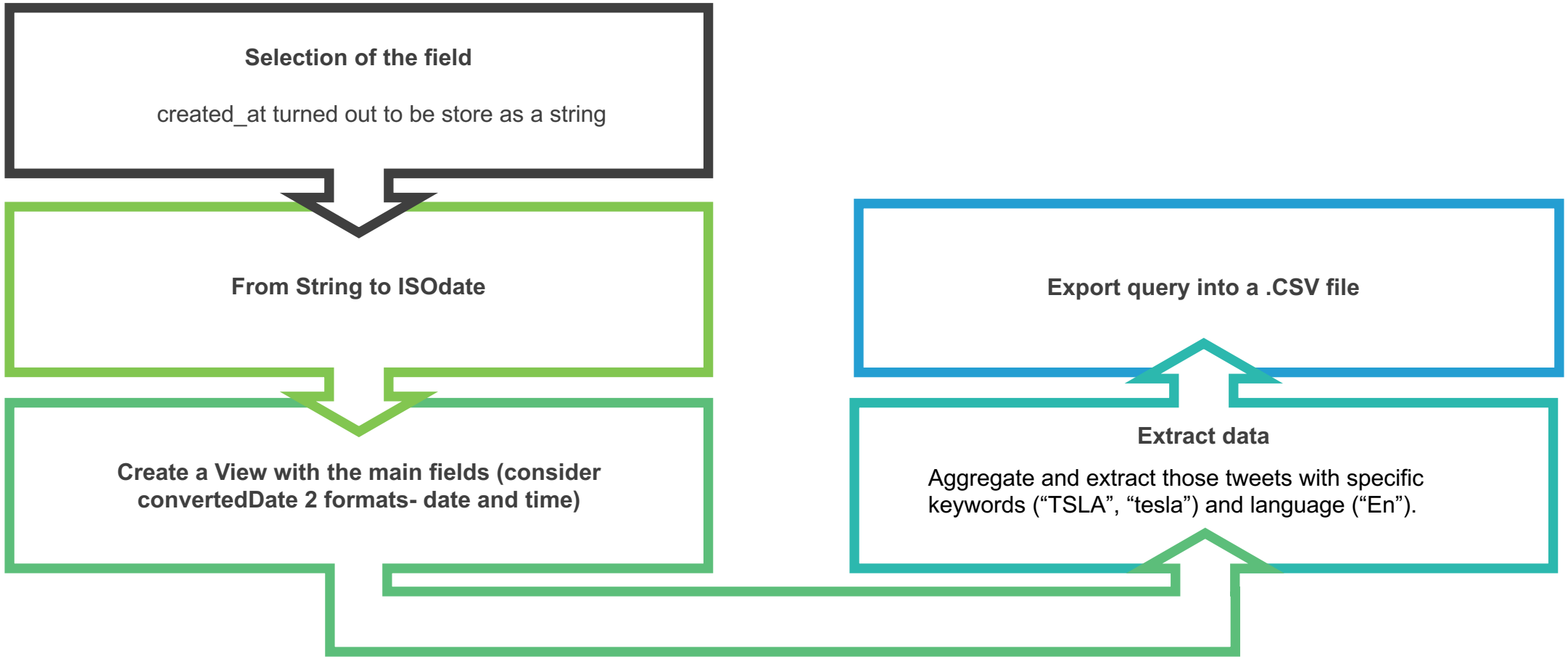
Data Processing



From .JSON files to .CSV

https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

Data Processing



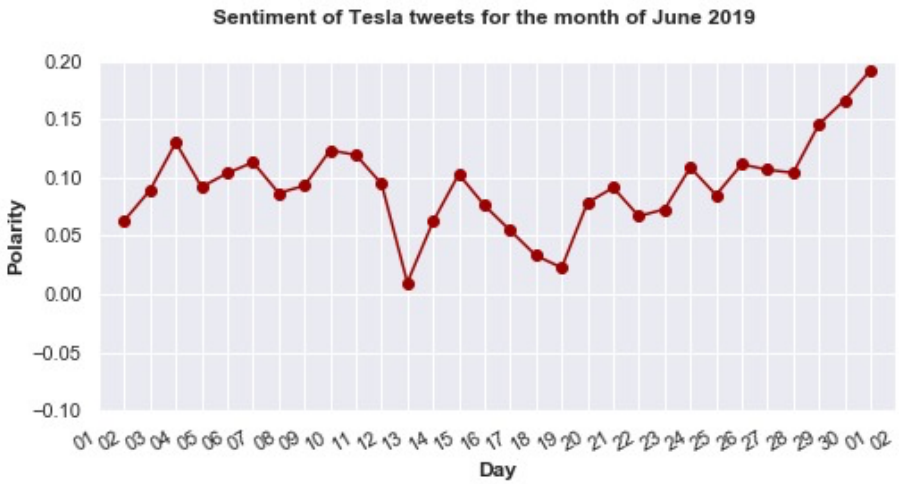
Sentiment Analysis

150.478.895 original June tweets

28.322.399 in English

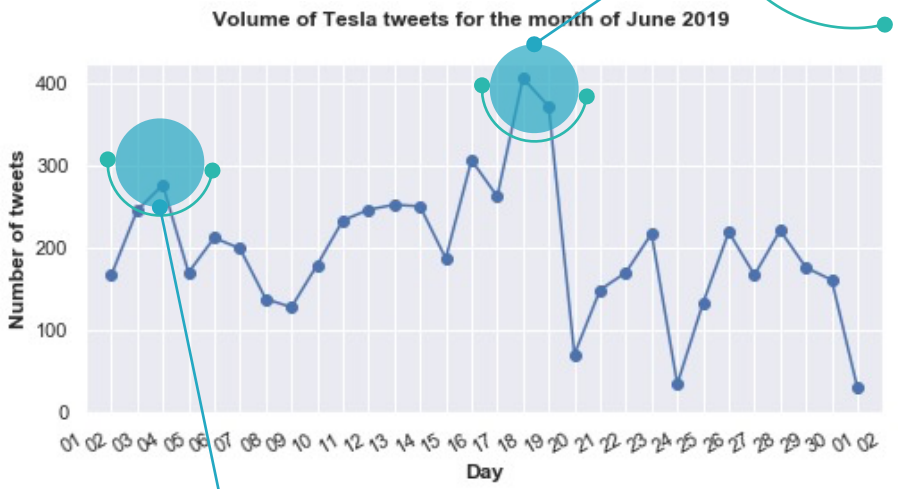
5.976 tweets containing “TSLA” or “TESLA”

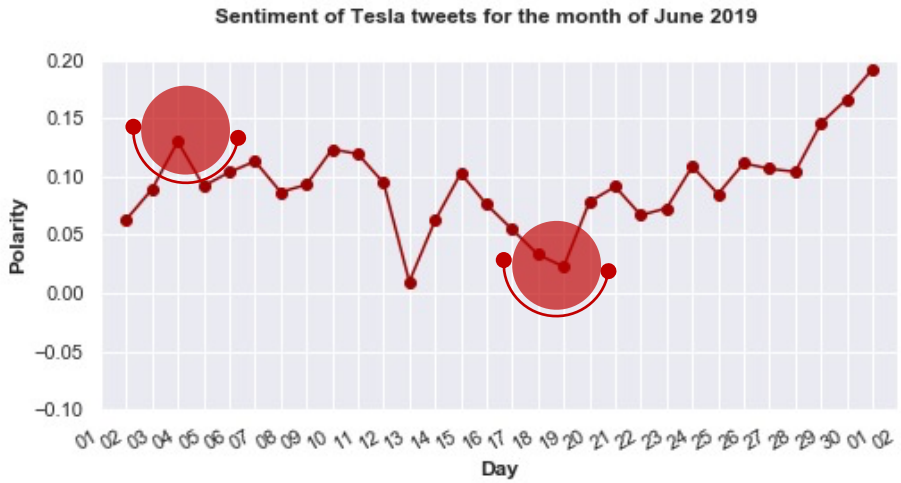
Cleaning Data + TextBlob



Sentiment Analysis

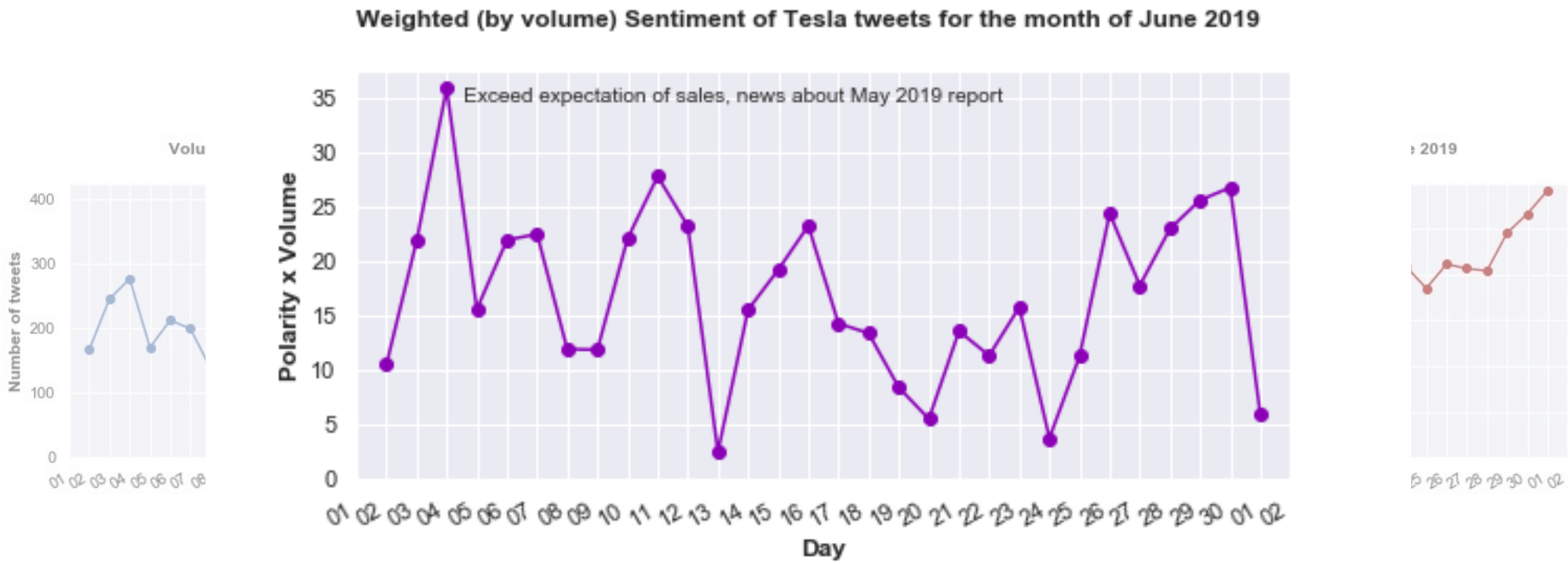
CEO Elon Musk: "Tesla is alive in spite of Eberhard, but he seeks credit constantly & fools give it him."





May's report
Sales Model 3 Rise, demand still not recovered
Doubts on Tesla's demand and growth prospects

Sentiment Analysis



Stock Price Data

Stock Price Data

Stock Price Prediction

MODELS:

- Naïve Model
- ARIMA Model
- SARIMA Model
- ARIMAX and SARIMAX Models

METRICS:

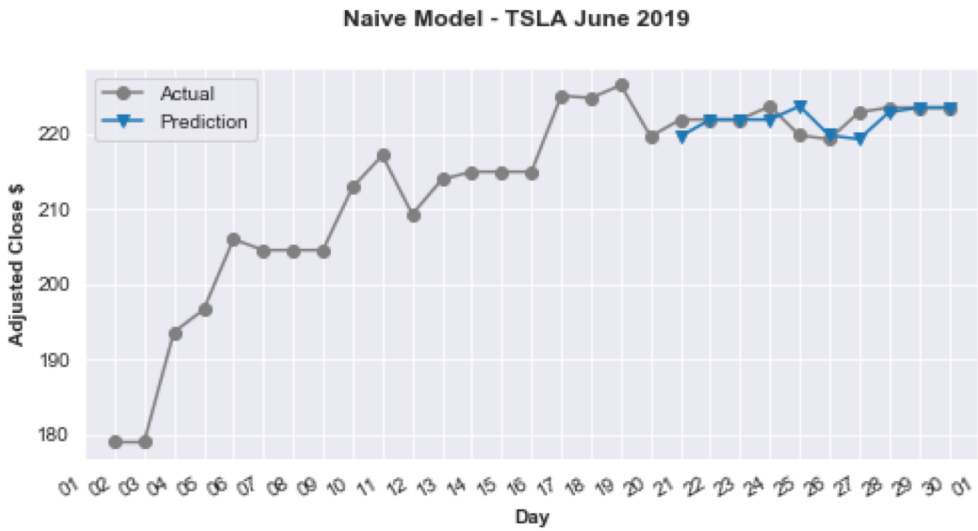
- Root Mean Squared (RMSE)
- Mean Absolute Percentage Error (MAPE)

Stock Price Prediction Models

- Naïve Model → $y(t+1) = y(t)$
- ARIMA Model
- SARIMA Model
- ARIMAX and SARIMAX Models

Test RMSE: 3.44 $

Test MAPE: 1.25%



Stock Price Prediction Models

- Naïve Model
- **ARIMA Model**
- SARIMA Model
- ARIMAX and SARIMAX Models

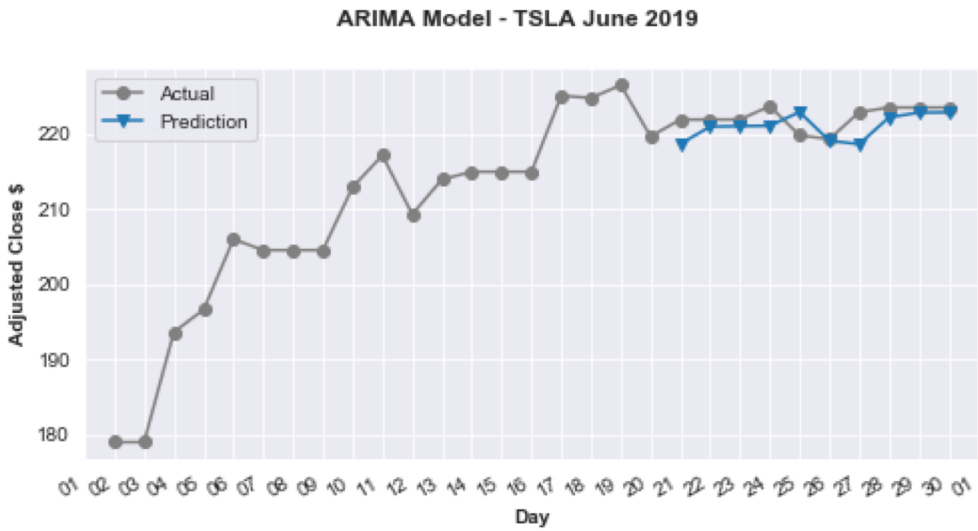
Auto-**R**egressive (p)

Integrated (d)

Moving **A**verage (q)

Test RMSE: 3.41 $

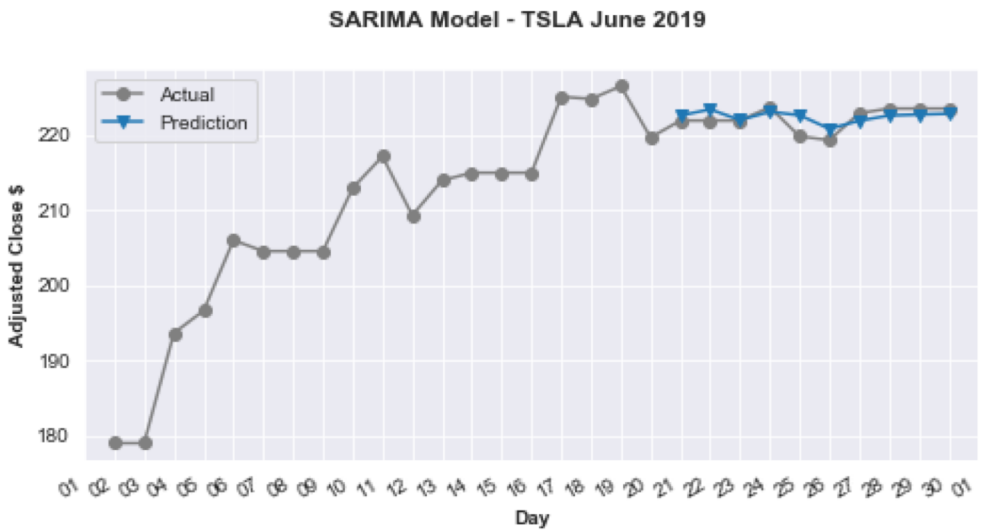
Test MAPE: 1.14%



Stock Price Prediction Models

- Naïve Model
- ARIMA Model
- SARIMA Model
- ARIMAX and SARIMAX Models

Seasonal ARIMA (p,d,q) $\mathbf{(P,D,Q)}_{\boldsymbol{m}}$



Test RMSE: 2.28 $

Test MAPE: 0.98 %

Stock Price Prediction Models

- Naïve Model
- ARIMA Model
- SARIMA Model
- ARIMAX and SARIMAX Models → (S)ARIMA + **Sentiment**

Low correlation (-0.05)

Conclusions

Stock price prediction: a very complex problem

• We considered only few data points •

• Sentiment analysis could be improved (SpaCy NLP framework) •

• More data sources (financial news) •

Thank You