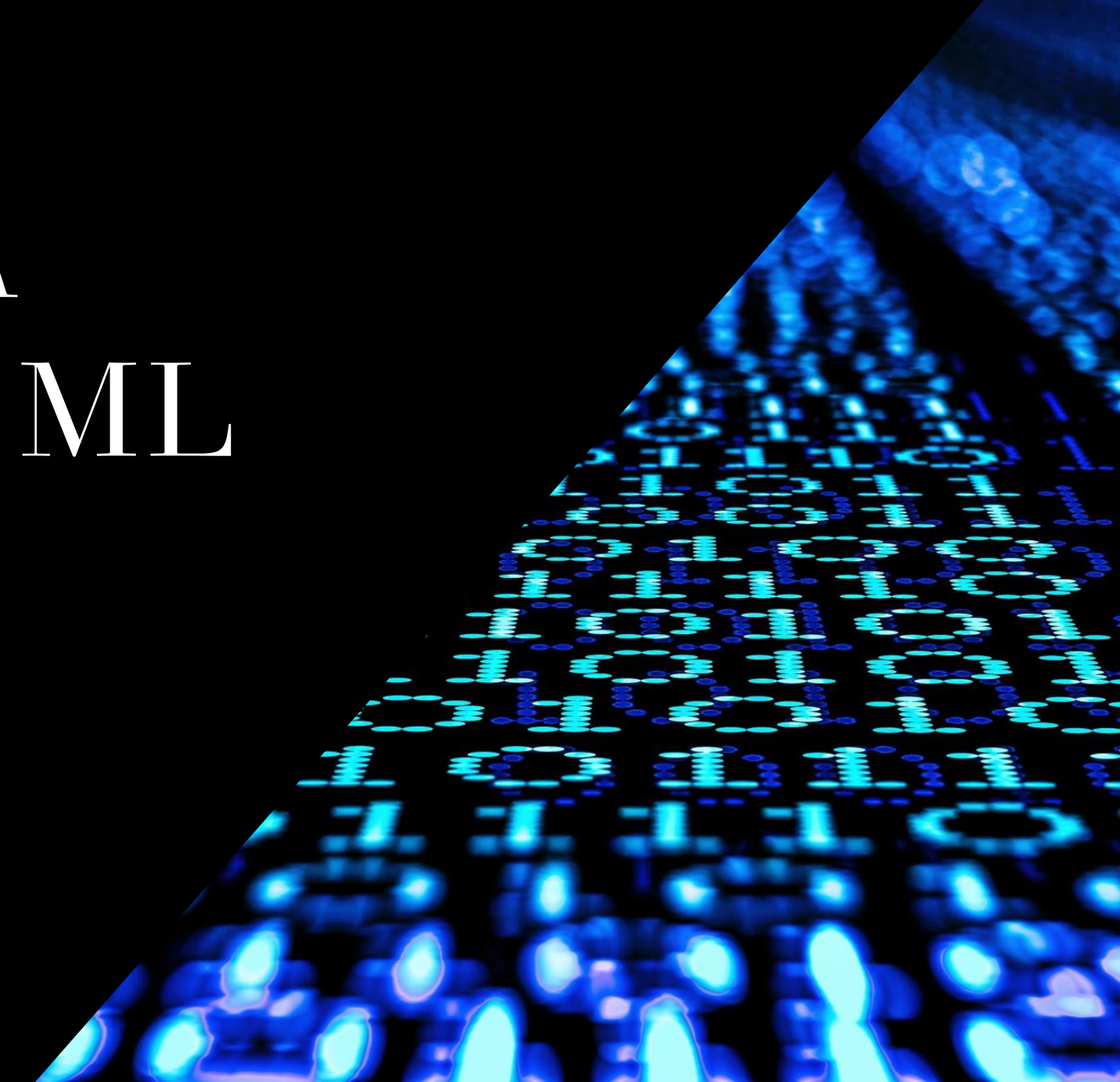
BIG DATA PYSPARK ML

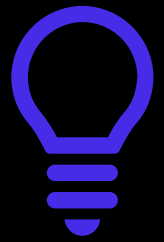


Aurélien Visentin

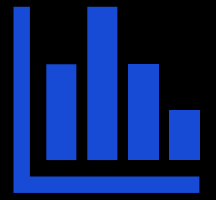
Eliot Leclair

Sommaire

Exploration

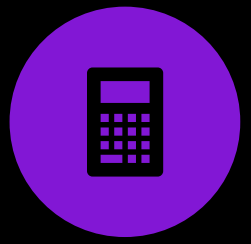


Insights



Regression

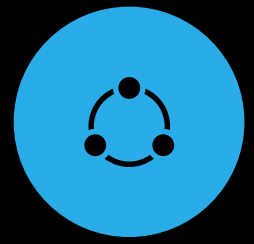
Exploration



MOYENNE MOBILE

RENDEMENT QUOTIDIEN

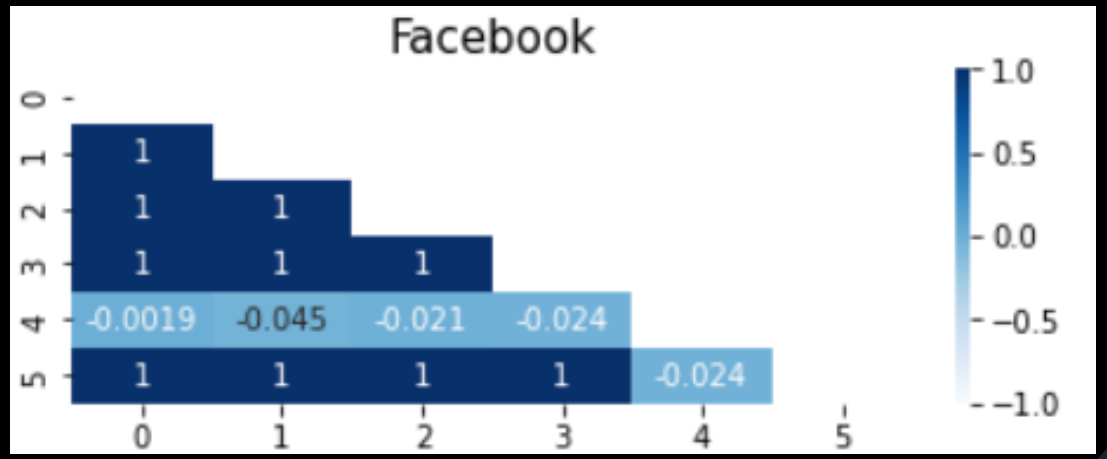
TAUX DE RETOUR SUR UNE PÉRIODE



CORRÉLATION

Exploration

```
The return rate of the amazon dataframe from the 2020-01-01 over a period of one month is: -27.3299560546875.

The return rate of the apple dataframe from the 2020-01-01 over a period of one month is: 5.879997253417969.

The return rate of the facebook dataframe from the 2020-01-01 over a period of one month is: -0.25.

The return rate of the google dataframe from the 2020-01-01 over a period of one month is: 88.469970703125.

The return rate of the microsoft dataframe from the 2020-01-01 over a period of one month is: 12.160003662109375.

The return rate of the tesla dataframe from the 2020-01-01 over a period of one month is: 42.11000061035156.

The return rate of the zoom dataframe from the 2020-01-01 over a period of one month is: 5.819999694824219.

------------ BEST RETURN RATE ------------
The best return rate from the 2020-01-01 over a 'month' period is: google with a rate of 88.469970703125.
```




```
========= Correlation between Amazon and Apple =========
+----------------+-------------------+--------------------+-------------------+--------------------+-------------------+-------------------+
|      Col_amazon|         High_apple|           Low_apple|         Open_apple|         Close_apple|       Volume_apple|    Adj Close_apple|
+----------------+-------------------+--------------------+-------------------+--------------------+-------------------+-------------------+
|     High_amazon| 0.9297874259390847|  0.9288404935316446|  0.929224429531747|  0.9289226688309318| 0.1614264488200611|  0.931117345769924|
|      Low_amazon| 0.9294814062048777|  0.9294547100635051| 0.9292104679328616|   0.929201231245745|0.14280092623787832| 0.9313750153531389|
|     Open_amazon| 0.9295864367635821|  0.9288996867205574|  0.929432373427345|  0.9286513439934997|0.15477270907314372| 0.9308299253600888|
|    Close_amazon| 0.9293612105453593|   0.928970764552425| 0.9287589351197002|  0.9292860084204524|0.15015419902613378| 0.9314678774869076|
|   Volume_amazon|0.07251513094870253|0.054936587225105346|0.06570150518170476|0.061806672508895474| 0.5340081166420992|0.06255731232524933|
|Adj Close_amazon| 0.9293612105453593|   0.928970764552425| 0.9287589351197002|  0.9292860084204524|0.15015419902613378| 0.9314678774869076|
+----------------+-------------------+--------------------+-------------------+--------------------+-------------------+-------------------+
```

Insights

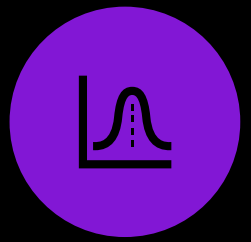
```
+------------+-----------------+---------------+
|company_name|investment_amount|expected_return|
+------------+-----------------+---------------+
|       apple|             10.0|           0.04|
|    facebook|             10.0|           0.06|
|      amazon|             10.0|            0.1|
|   microsoft|             10.0|           0.14|
+------------+-----------------+---------------+

We expect 8.5% of returns
```

Modern Portfolio Theory

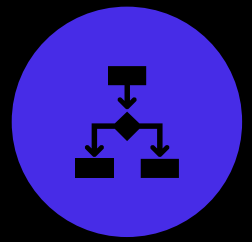
```
schema = StructType([
  StructField("message", StringType(), True),
  StructField("result", ArrayType(
    StructType([
      StructField("Column1", IntegerType()),
      StructField("Column2", StringType())
    ])
  ))
])
```

REST API Call

Régression



LINEAR REGRESSION



DECISION TREE

RANDOM FOREST

TUNED RANDOM FOREST

Régression


Amazon: linear_regressor
train(Adj Close)
train(linear_regressor)
test(Adj Close)
test(linear_regressor)
3500
3000
2500
2000
1500
1000
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01

Régression

Limitations de l'algorithme