INSTITUT POLYTECHNIQUE DE PARIS

MAP670G - Data Stream

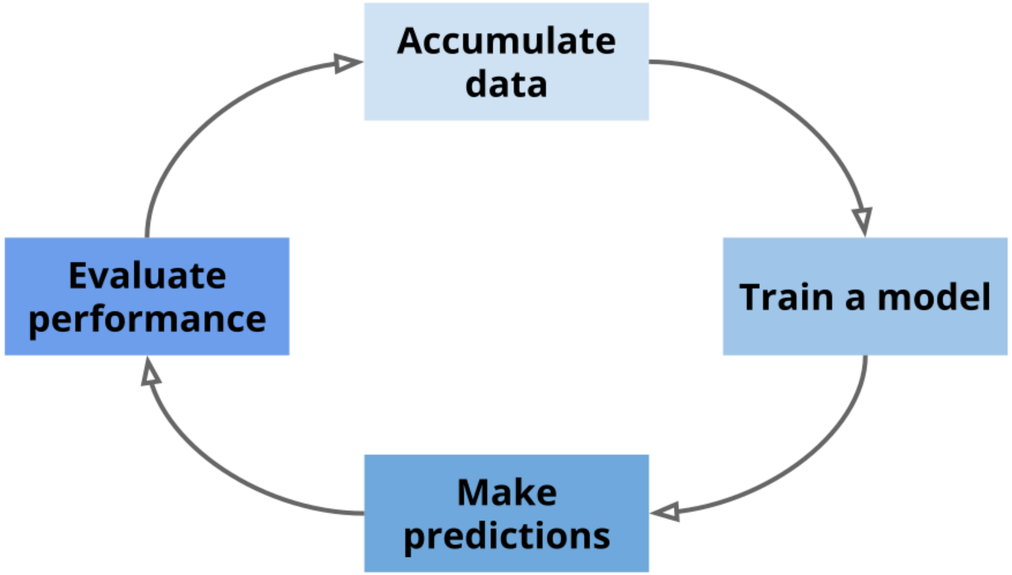
# Real-time streaming application with Kafka

## Collect trading data using Yahoo finance API and use online regression to predict markets stocks

Laurine Burgard-Lotz, Emma Milliot, Margaux Boscary

# Introduction : Explanation about the problem

Machine Learning



Online Learning

- **Google** for the USA
- **BNP Paribas** for France
- **Alibaba** for China

# Outline

COLLECTING DATA

STREAMING USING KAFKA

TRAINING THE MODEL

MAKING THE PREDICTION IN REAL TIME

ANALYSIS OF THE RESULTS

# APIs for stock market data

Yfinance 0.1.70

Several interval of time for the data (shortest = 1 minute), easy of use (python librairy), access to *Open, High, Low, Close, Volume, Dividends, Stock Splits*

yahoo! finance

Issues : $$$ + no close price

ALPHA VANTAGE

Issues : not in real time

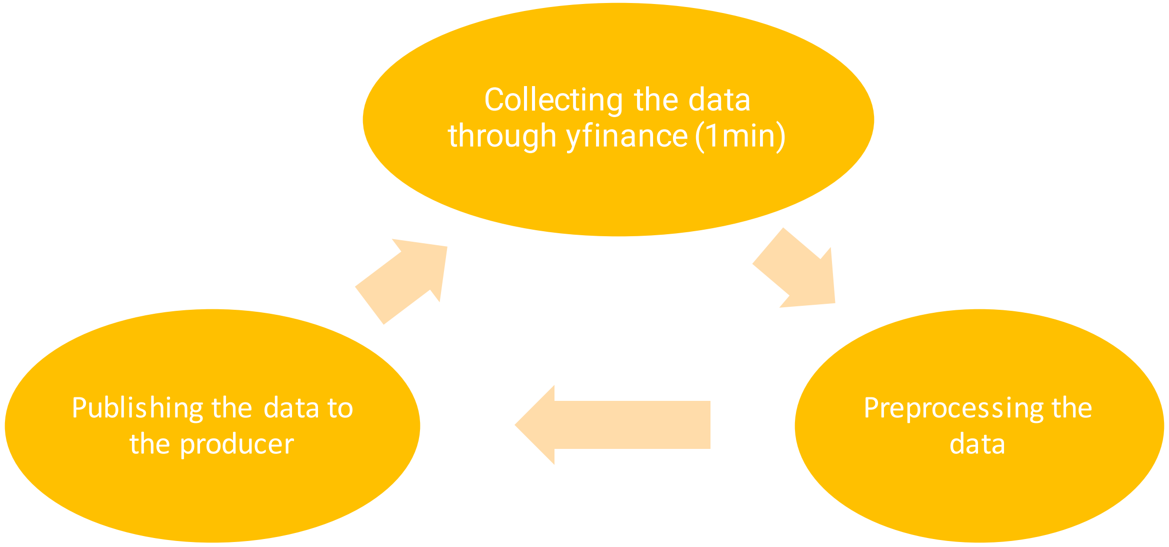
# Kafka

Open-source distributed event streaming platform

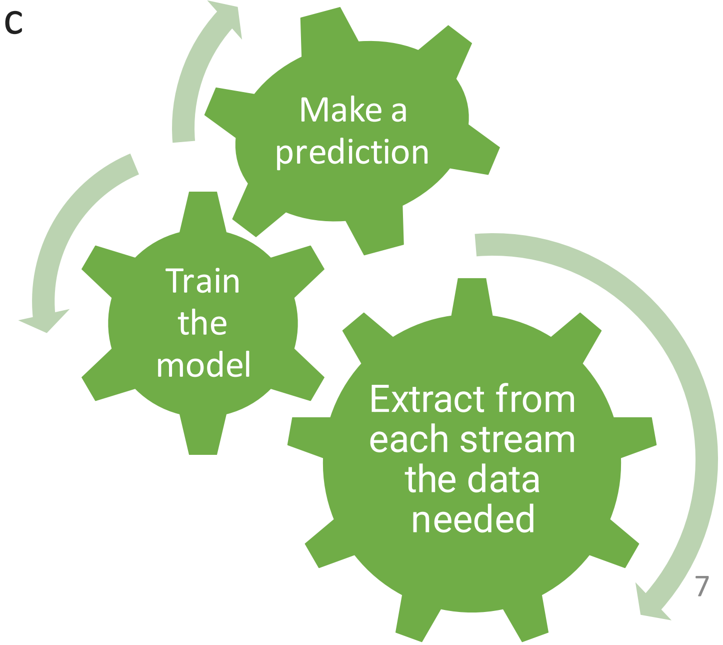
| **Allows to**: | |
|---|---|
| | Write and read streams of events, including continuous import/export of the data from other systems. |
| | Store streams durably and reliably for as long as it is needed. |
| | Process streams as they occur or retrospectively. |

## Our use case :

**The producer :** automate the retrieval of financial data


Collecting the data through yfinance (1min)
Preprocessing the data
Publishing the data to the producer

**The consumer:** access to the data stocked into the Kafka topic



# The **online learning model** : training & fine-tuning

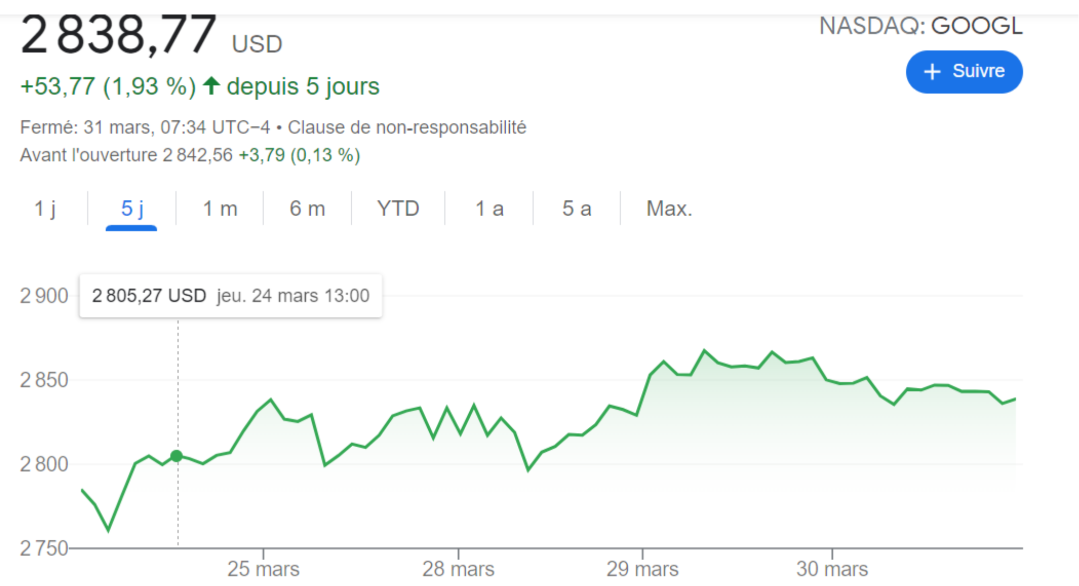
- Linear regression
- Stochastic Gradient Tree
- Hoeffding Tree
- Hoeffding Adaptive Tree.

**Let's see it on the notebook**

# Results

| Country | Stock | Model | MAE Online | MSE Online | RMSE Online |
|---|---|---|---|---|---|
| USA | Google | Linear regression | 31,9 | 1,24E+04 | 111 |
| | | SGT Regressor | 28,4 | 6,40E+04 | 253 |
| | | MLP | 2,81E+03 | 7,92E+06 | 2,81E+03 |
| | | Hoeffding Tree | 3,97 | 3,74E+01 | 6,12 |
| | | Hoeffding Adaptative Tree | 2,92 | 2,38E+01 | 4,88 |
| France | BNP Paribas | Linear regression | 0,184 | 2,31E+00 | 1,52 |
| | | SGT Regressor | 0,612 | 1,61E+01 | 4,01 |
| | | MLP | 51,3 | 2,64E+03 | 51,3 |
| | | Hoeffding Tree | 0,0354 | 0,00243 | 0,0493 |
| | | Hoeffding Adaptative Tree | 0,0717 | 3,30E-01 | 0,574 |
| China | Alibaba | Linear regression | 0,594 | 1,44E+01 | 3,79 |
| | | SGT Regressor | 1,36 | 9,79E+01 | 9,89E+00 |
| | | MLP | 115 | 1,31E+04 | 115 |
| | | Hoeffding Tree | 0,76 | 1,28E+00 | 1,13 |
| | | Hoeffding Adaptative Tree | 0,645 | 4,20E+00 | 2,05 |

## Google



2 838,77 USD

NASDAQ: GOOGL

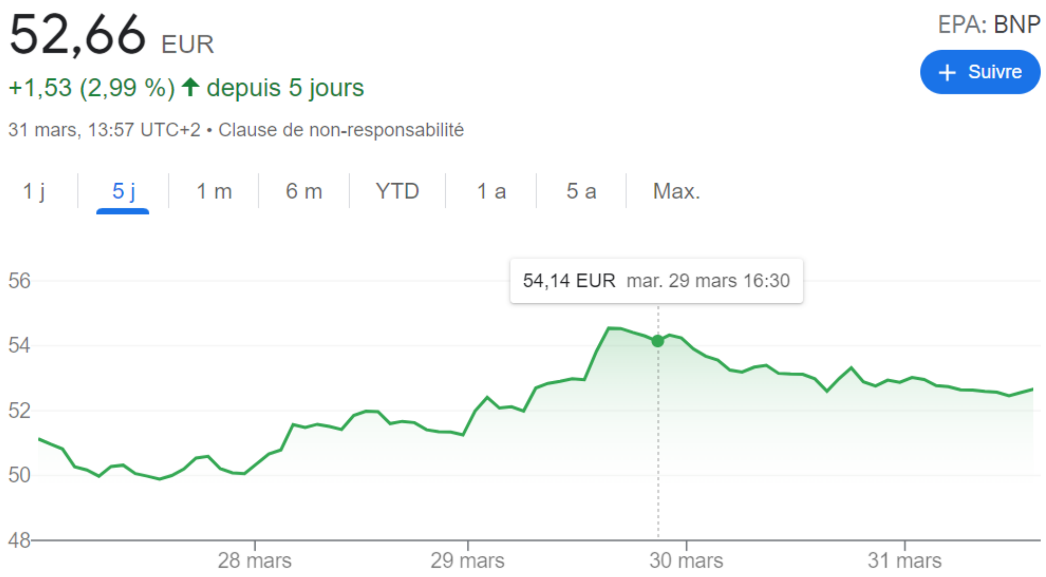
+53,77 (1,93 %) ↑ depuis 5 jours

Fermé: 31 mars, 07:34 UTC−4 • Clause de non-responsabilité

Avant l'ouverture 2 842,56 +3,79 (0,13 %)

2 805,27 USD jeu. 24 mars 13:00

## BNP



52,66 EUR

EPA: BNP

+1,53 (2,99 %) ↑ depuis 5 jours

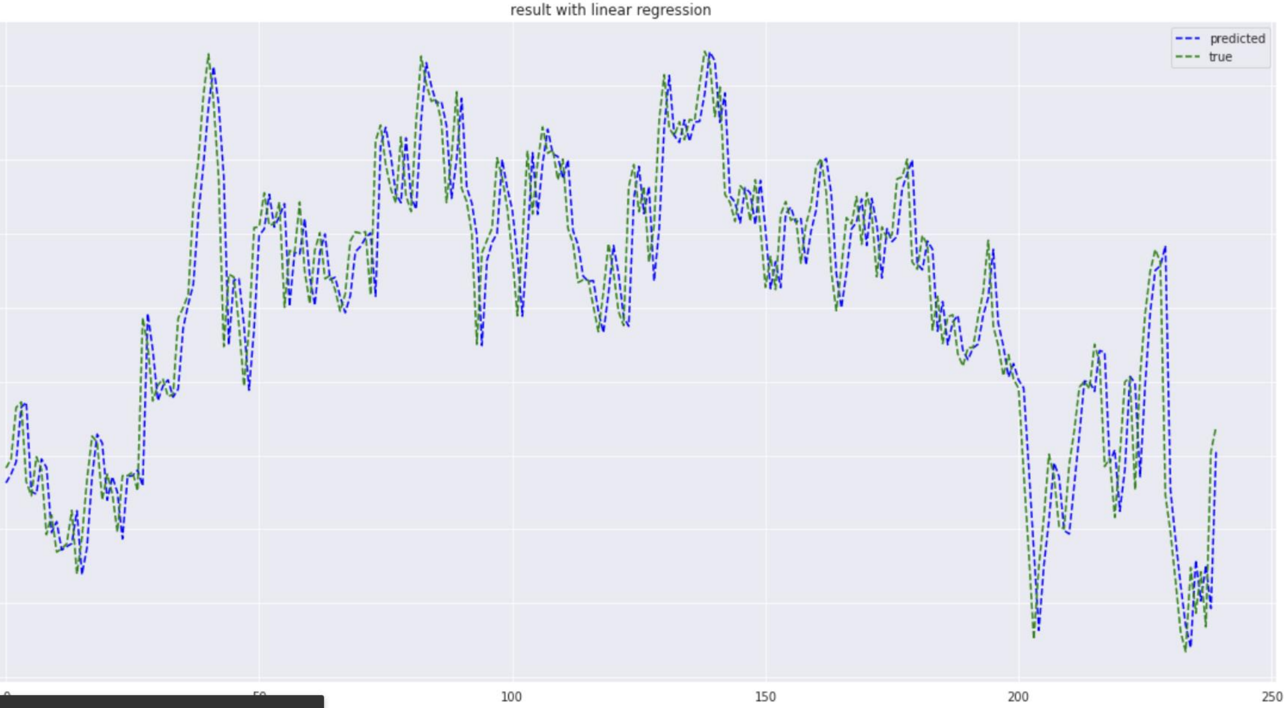
31 mars, 13:57 UTC+2 • Clause de non-responsabilité

54,14 EUR mar. 29 mars 16:30

# Démo

# Comparison with batch models

In the 3 differents stocks, the linear regression is by far the best.
The predicted curve is shifted by one minute.


result with linear regression

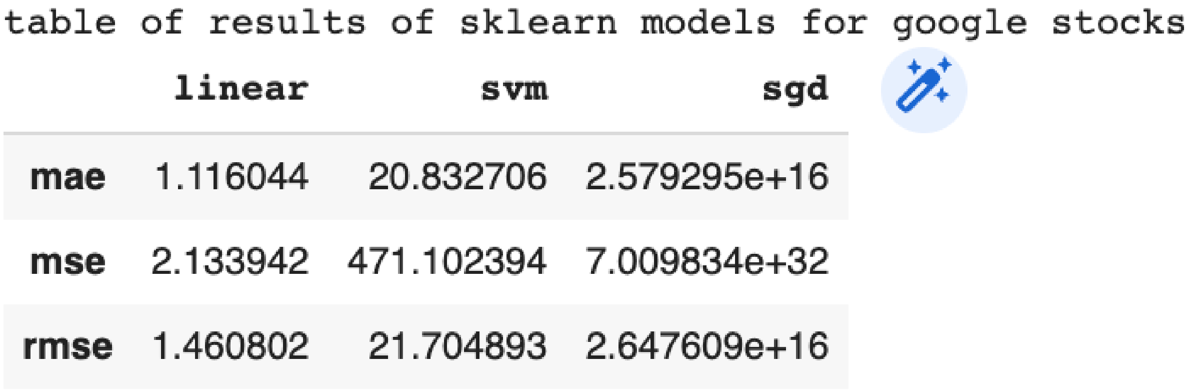


table of results of sklearn models for google stocks

| | linear | svm | sgd |
|---|---|---|---|
| **mae** | 1.116044 | 20.832706 | 2.579295e+16 |
| **mse** | 2.133942 | 471.102394 | 7.009834e+32 |
| **rmse** | 1.460802 | 21.704893 | 2.647609e+16 |

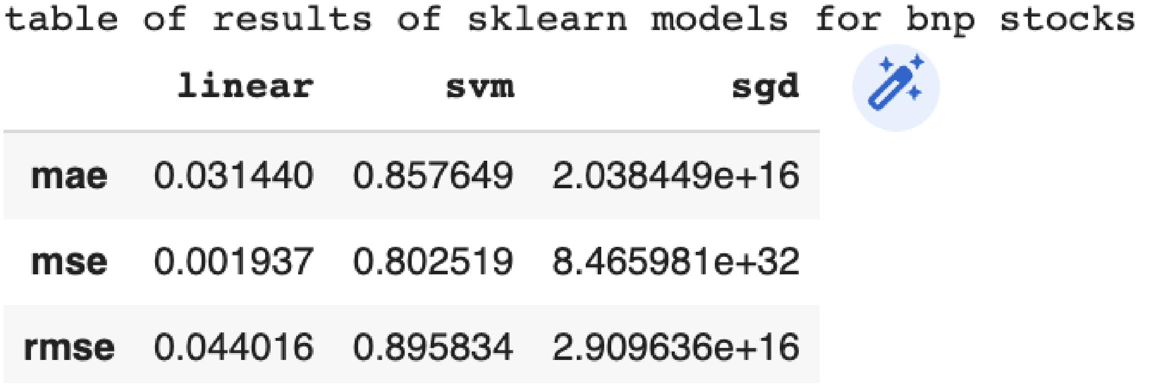


table of results of sklearn models for bnp stocks

| | linear | svm | sgd |
|---|---|---|---|
| **mae** | 0.031440 | 0.857649 | 2.038449e+16 |
| **mse** | 0.001937 | 0.802519 | 8.465981e+32 |
| **rmse** | 0.044016 | 0.895834 | 2.909636e+16 |

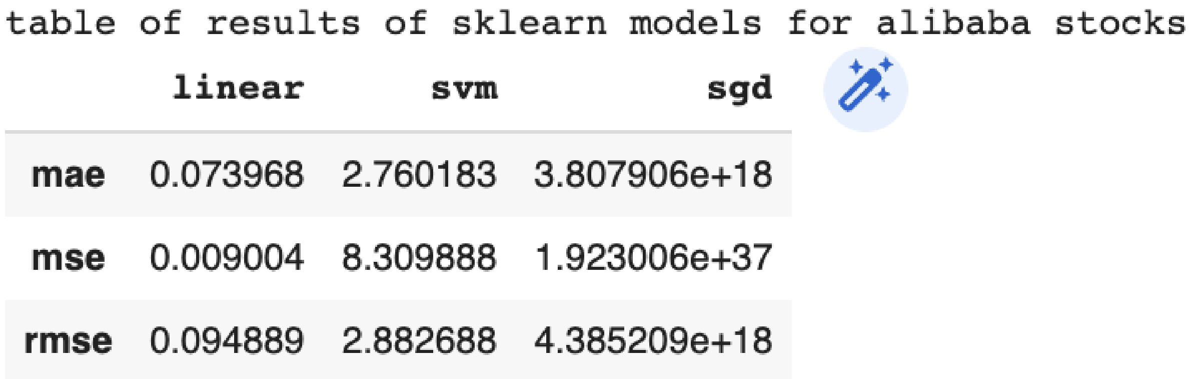


table of results of sklearn models for alibaba stocks

| | linear | svm | sgd |
|---|---|---|---|
| **mae** | 0.073968 | 2.760183 | 3.807906e+18 |
| **mse** | 0.009004 | 8.309888 | 1.923006e+37 |
| **rmse** | 0.094889 | 2.882688 | 4.385209e+18 |

# Conclusion

## Online learning

- Lack of training data
- Adapted to change

## Batch learning

- Faster training
- Better results