Hackathon SDD 2023

Hackathon Partners and Subjects

MERCATOR OCEAN INTERNATIONAL

1. **Ocean Eddy Identification**

Capgemini invent

2. **Building energy consumption estimation**

METEO FRANCE

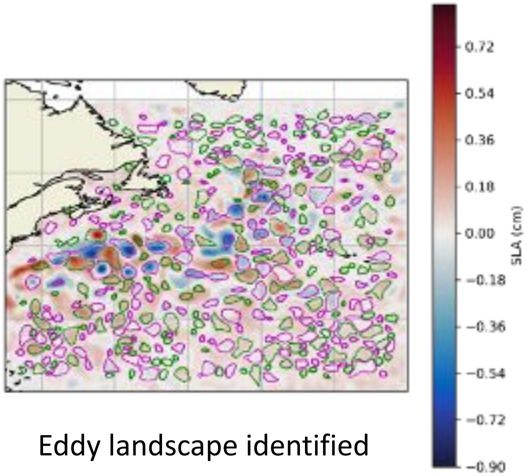
3. **Daily Rainfall Forecasting**

AIRBUS DEFENCE & SPACE

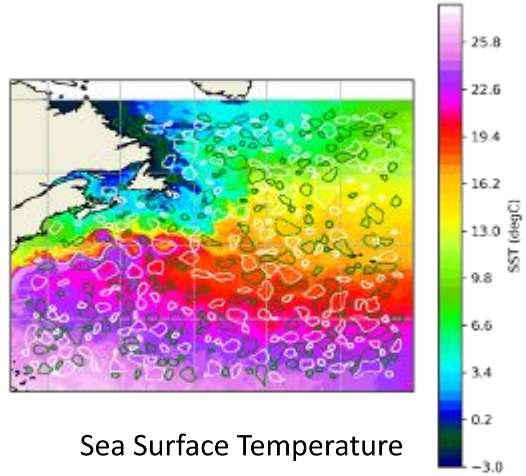
4. **Satellite Maneuver Detection**

Subject 1: Ocean Eddy Identification

Responsable : Daria Botvynko, Simon van Gennip



Eddy landscape identified
With Sea Level Anomaly



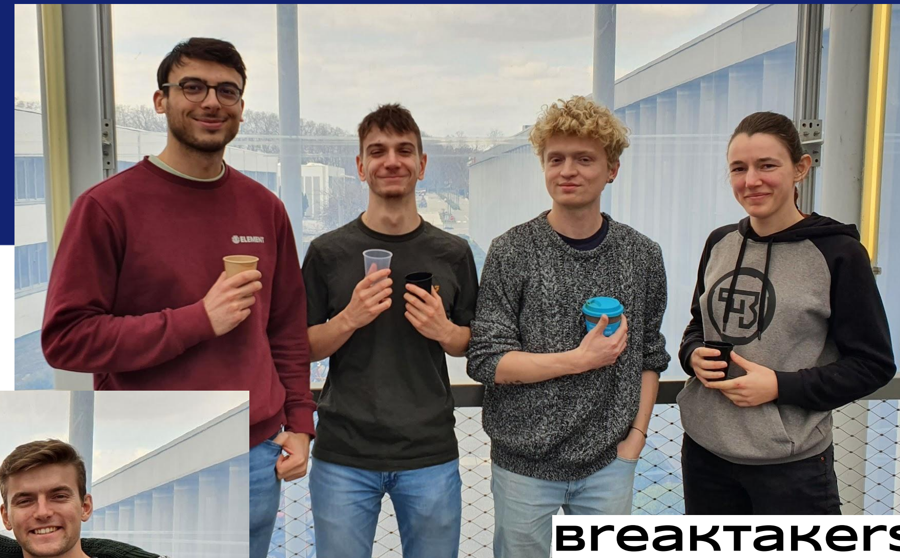
Sea Surface Temperature
(SST)

Oceanic Eddies are vortices of the order of ~10km in horizontal scale, whose signature is clearly visible in satellite products and well reproduced in ocean models. Sea Level Anomalies (SLA) are used for detecting such object by means of numerical techniques, yet such approach lacks accuracy in detection, namely because Sea Level Anomaly contains errors. Eddy signature is also visible in other variables such as Sea surface Temperature (SST) that do not suffer from such limitation.

The objective here is to **develop a Deep Learning approach to identify eddies using SST, SLA and the ocean velocity field.** For this students will have a dataset consisting of SLA, SST, and velocity components images which are slightly distorted relative to reality, together with labels (the eddy contours, see figure) obtained directly from reality.

MERCATOR OCEAN INTERNATIONAL


L'EQUIPE 21


Breaktakers


Les menteurs


Clic Clic Pan Pan


Sleepy Hamilton

Subject 1 Results

What is the scoring metric ?

Who is the winner ?

#	Team	Members	Score	Entries	Last	Join
1	Théo Rousseaux		120.75269	8	1h	
2	Break Takers		64.07039	9	40m	
3	L'ékip 21		0.81047	5	18m	
4	Sleepy Hamilton		0.67083	10	5m	

Subject 1 Results

What is the scoring metric ?

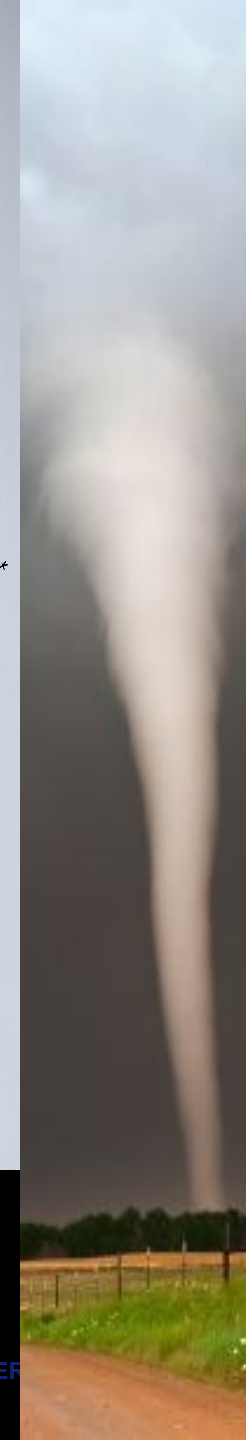
ClicClicPanPamjesaispasquoi

Who is the winner ?

#	Team	Members	Score	Entries	Last	Join
1	Théo Rousseaux		120.75269	8	1h	
2			64.07039	9	40m	
3			0.81047	5	18m	
4			0.67083	10	5m	

Eddies Detection

by the Clic Clip Pan Pan Team



Planning

Lundi : Mise en place

- visualisation des données, compréhension du problème
- Elaboration de notre stratégie : problème de segmentation, choix d'une architecture CNN U-net et assignation des tâches
- Pre-processing, classe de notre modèle, formalisation du dataset
- Décision de gestion de la terre sur les images avec un masque

Mardi : Améliorations

- premiers essais d'entraînement pour une époque
- fonction de lissage pour éviter les discontinuités au niveau des bords de terre
- Data Augmentation
- implémentation d'une meilleure fonction loss ne prenant pas en compte les pixels de Terre
- premier entraînement avec des résultats satisfaisants

Mercredi : Finalisation

- correction des différents problèmes : loss, augmentation des données…
- lancement de plusieurs entraînements et optimisation des hyper-paramètres

Architecture de notre modèle

- Utilisation d'une architecture de **Unet** avec padding pour obtenir en sortie la même dimension que l'entrée
- 4 channels en inputs et 3 en outputs
- **Cross Entropy Loss** utilisée sur l'image de sortie masquée (pour enlever la terre)
- **F1 score** sur chaque classes en **one-vs-all** utilisée comme métrique d'évaluation

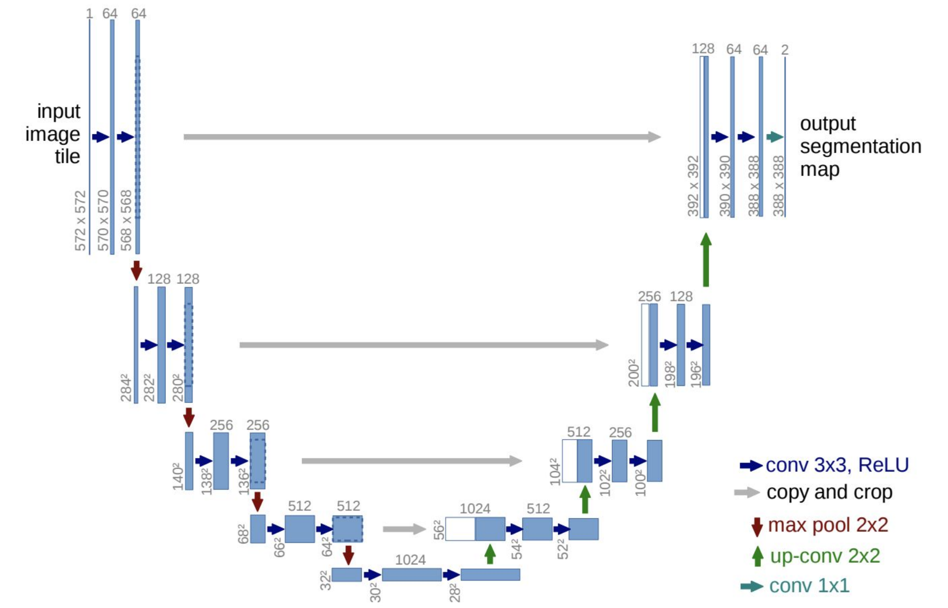
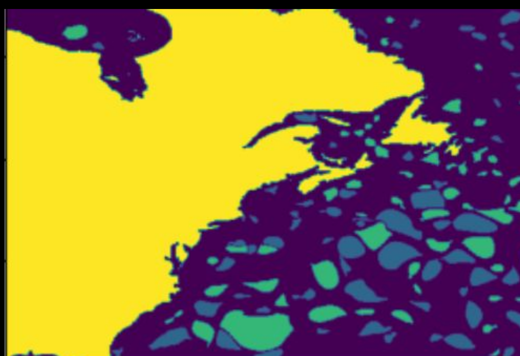


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

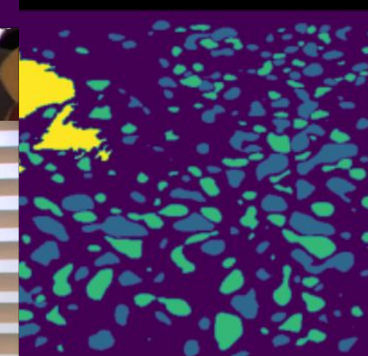
<u>Preprocessing :</u>

- Normalisation, lissage des bords de terre
- Augmentation : rotation 180°, bruitage des données

Résultats



Ground truth



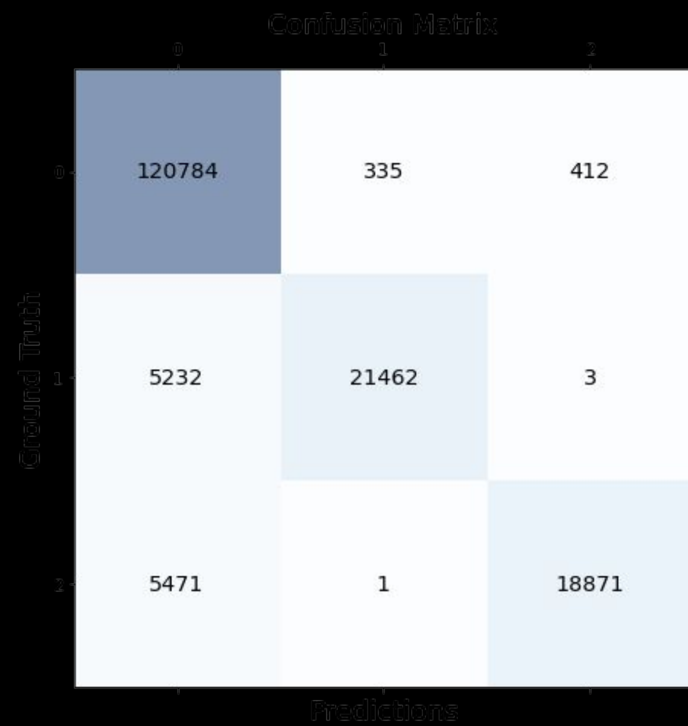
tput

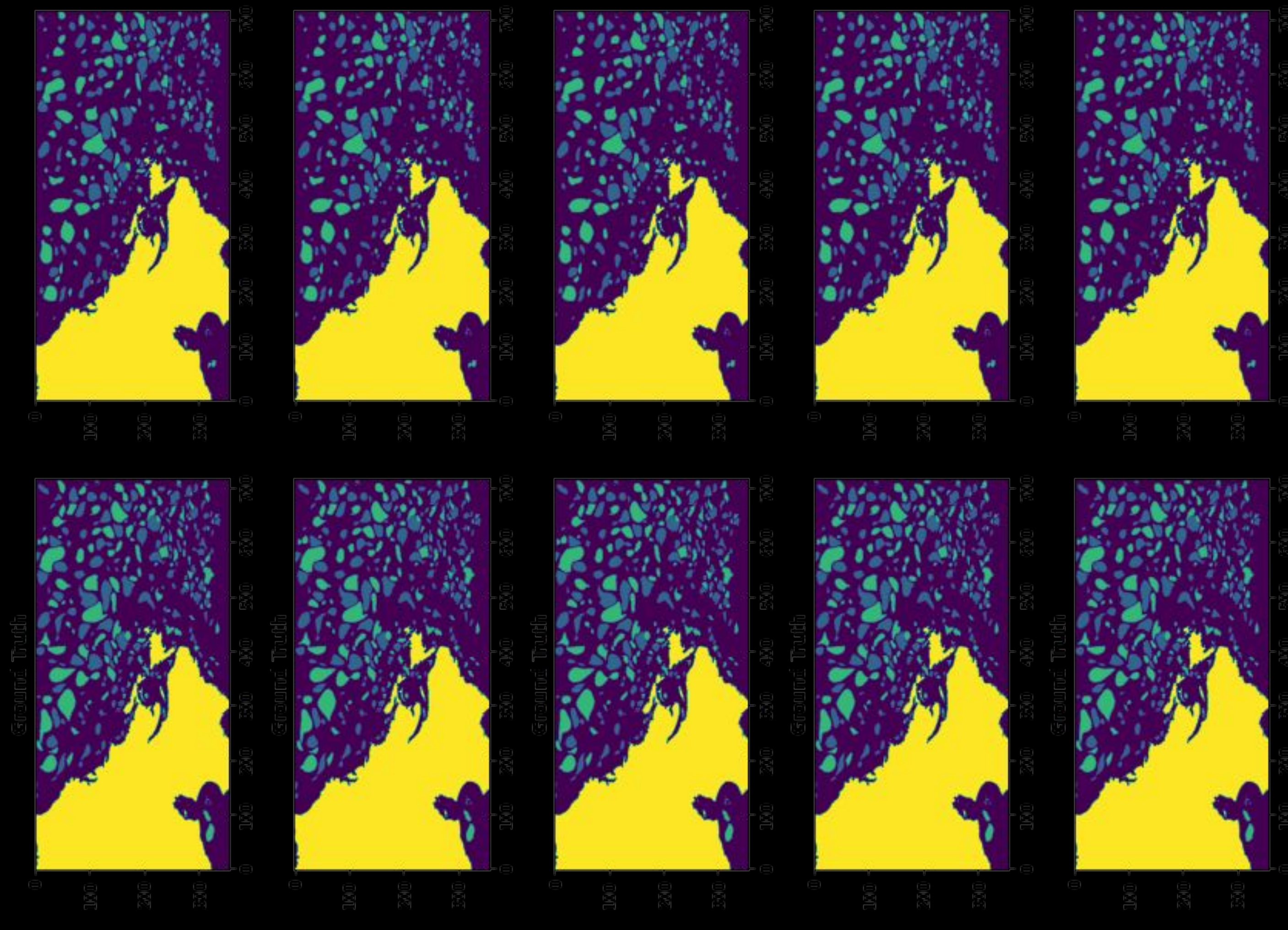
Résultats

Training of 45 epochs

F1 Score
0.995
0.885
0.865

Matrice de confusion

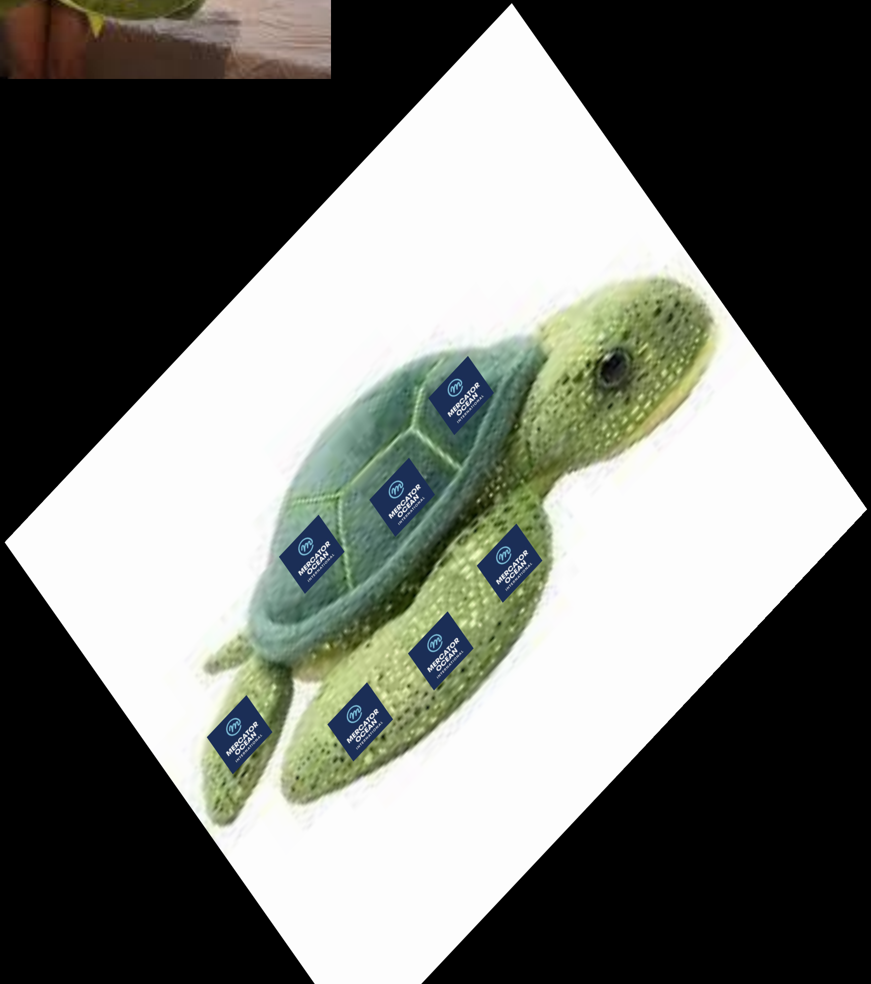




Scores Kaggle

result10.csv Complete · Florian Thomas · 22m ago · Modèle avec moins d'épochs (loss minimale)	0.84265
result_V4_final.csv Complete · Florian Thomas · 1h ago · Même modèle mais avec plus d'épochs (loss plus importante mais amélioration des F1 Score 1vsAll)	0.85996
result1.csv Complete · Florian Thomas · 2h ago · Il y'a méprise, j'ai upload le mauvais modèle	0.83472
result1.csv Complete · Florian Thomas · 2h ago · Quelques épochs de plus pour la forme	0.79697
result1.csv Complete · Florian Thomas · 3h ago · Détection un peu agressive, on verra bien	0.79697
result2.csv Complete · Florian Thomas · 5h ago · On sait jamais	0.81406

Where goodiiiiiiiiiieeeees :?




Au revoir.

Subject 2: Energy Consumption Estimation

Responsable : Lucas Lima Lopes, Louis Melliorat

CONTEXT:

The last years mark **a sharp acceleration in the energy renovation policy**. In the field of public buildings, local authorities will be at the heart of this dynamic. Thankfully, significant investments are being made to **improve building efficiencies** to reduce costs and emissions. The question is, **are the improvements working?** That's where you come in.

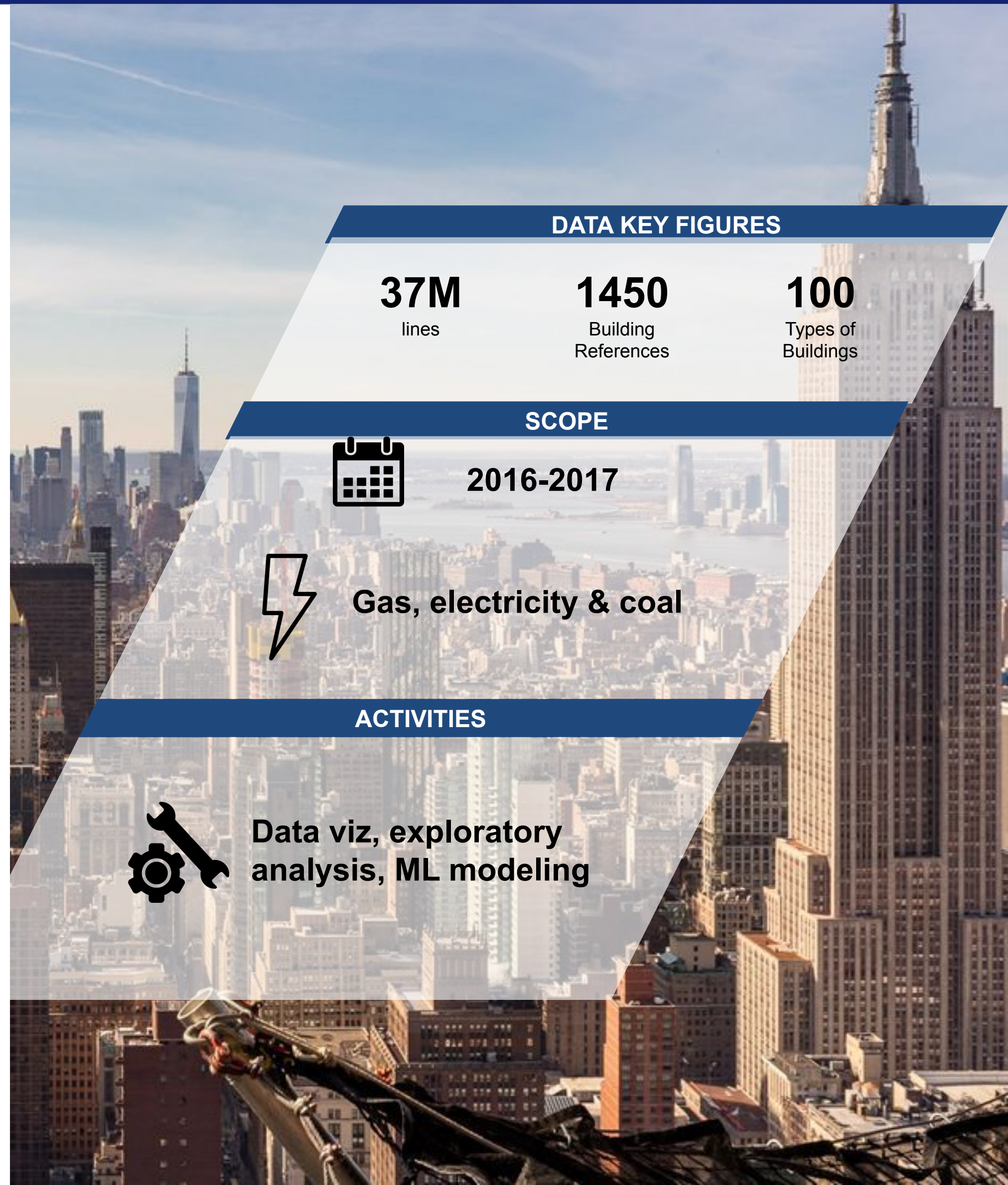
YOUR ROLE:

You are a **Data Science innovative Startup** that is solicited by **Public Services** to help them control their investments, monitor their building energy consumption and to recommend them the **best strategy to renovate their real estate**. You have three days to be able to build a strong recommendation to your client

YOUR MISSION:

In those three days, you will be able to build an **energy optimization strategy** by:

1) Treat the data, define the best **metrics** and **visualizations** to help your client **quantify its real estate energy consumption**
2) **Produce and train a model** to predict the energy consumption of a public building
3) **Define and present your strategy** to the client



Capgemini invent

Subject 2: Energy Consumption Estimation

THE DATASET

Buildings metadata

Surface, usage, latitude, etc

Historical data

Daily historic of energy consumption, source, etc

Weather data

Conditions at the site and time of the consumption measurement

Predict

Consumption

Predict the variable meter reading for each row identified by building id, timestamp and energy source

	timestamp	building_id	meter	site_id	primary_use	sub_primary_use	square_feet	lat	lng	air_temperature	cloud_coverage
0	2016-05-21	0	0	0	Education	Research	7432.0	28.52	-81.4	26.370833	5.958334
1	2016-05-22	0	0	0	Education	Research	7432.0	28.52	-81.4	26.554167	3.000000
2	2016-05-23	0	0	0	Education	Research	7432.0	28.52	-81.4	25.516667	4.791666
3	2016-05-24	0	0	0	Education	Research	7432.0	28.52	-81.4	25.262500	3.541667
4	2016-05-25	0	0	0	Education	Research	7432.0	28.52	-81.4	25.354167	2.583333

…

meter_reading
550.14300
506.46002
735.80100
221.83200
211.59400


supacafewpop


wizardly turing


Les trois olympiques


Habibi de Favela

Subject 2 Results

MODÈLE

#	△	Team	Members	Score	Entries	Last	Solution
1	—	Wizardly Turing		0.89051	32	5h	
2	—	Habibi de Favela		1.04043	25	5h	
		sample_.csv		1.08028			
3	—	Les trois olympiques		1.16313	41	4h	
4	—	pierrette_sup		1.27014	9	7h	
5	—	JohnnyDEP	A	1.40678	28	4h	

PRÉSENTATION

1. Habibi de Favela
2. Wizardly Turing
3. Les trois olympiques
4. JohnnyDEP

Subject 2 Winner Model/General

Prévoir la consommation énergétique des bâtiments

1. Détection d'outliers

Suppression des bâtiments avec :
- Une **consommation d'énergie trop élevée**
- Une **consommation d'énergie nulle**

2. Feature Engineering

OneHot Encoding des variables catégorielles

Sélection des features pertinentes :
- Les données météo les plus corrélées avec la consommation
- Latitudes et Longitudes

Hybridation de features pour prendre en compte :
- La saisonnalité annuelle,
- Le rythme hebdomadaire (semaine vs. week-end),
- L'influence de la météo des jours précédents

3. Modèle : Gradient Boosting

Modèle de **gradient boosting** (Random Forest optimisé)

Concaténation de **4 sous-modèles** prédisant chacun une source d'énergie

Hyperparameters tuning avec **RandomizedSearch**

4. Résultats de prédiction

MAE pour la prédiction d'énergie :
- Electricité : 34,1%
- Eau froide : 11,1%
- Vapeur : 12,7%
- Eau chaude : 1,8%

Paramètres importants : surface des bâtiments, température de l'air, jour de la semaine vs. week-end, latitude et longitude

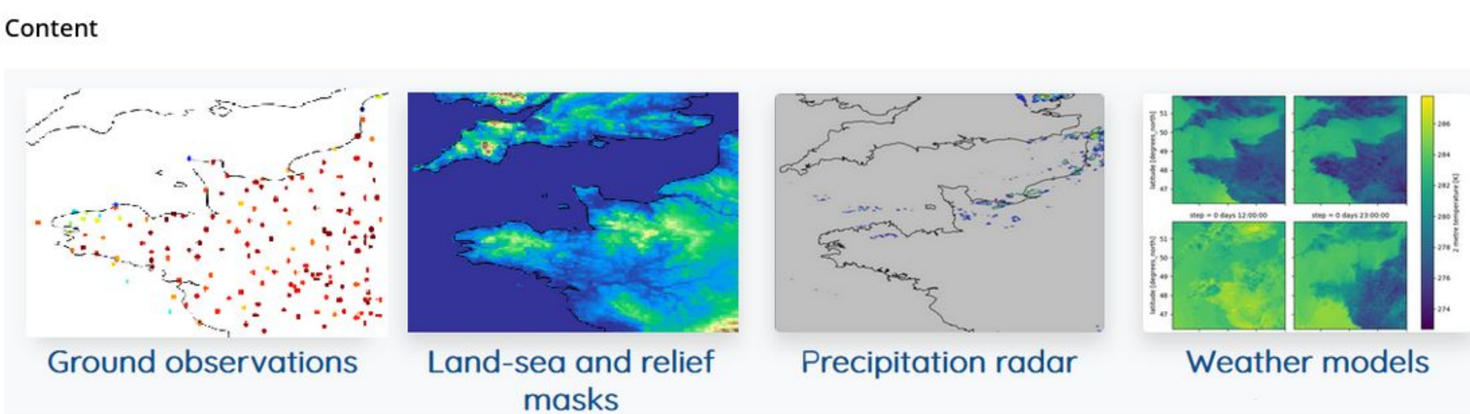
Subject 3: Daily Rainfall Forecasting

Responsable : Léa Berthomier

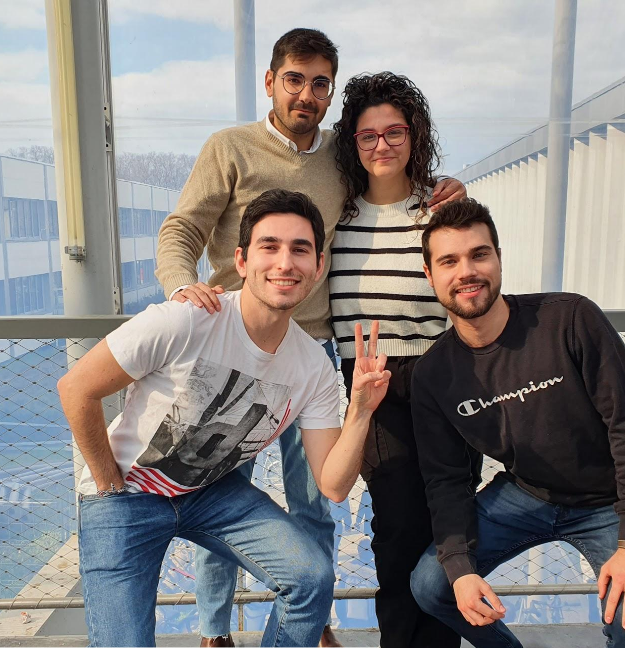
Forecasting the daily rainfall can prevent you from coming back from work soaked from head to toe, but more importantly, it can help anticipating extreme events such as floods or hurricanes.

Your mission, should you choose to accept it, will be to forecast the rainfall over the next 24h using data from the previous day : ground observations (temperature, wind, pressure, rainfall…) and numerical weather forecasts from METEO-FRANCE. (These forecasts are based on the equations of the physics of the atmosphere).
You will have to learn from the past errors of the numerical weather models to forecast the rainfall on several ground stations.

The data comes from MeteoNet, an open weather dataset for AI : https://www.kaggle.com/katerpillar/meteonet
It spans the North-West quarter of France from 2016 to 2019.



METEO FRANCE



Jolly Darwin

Hackatruite

Busy Maxwell

Funny Shockley

Inzeboite

Subject 3 Results

#	△	Team	Members	Score	Entries	Last	Join
1	▲1	Hackatruite		28.47519	33	5m	
2	▲1	InZeBoite		28.76340	22	19m	
3	▼2	Jolly Darwin		29.37246	44	22m	
4	—	Funny Shockley		32.55429	47	4m	
5	—	Des 8A comme ça 👍		34.66995	3	2d	
6	—	Busy Maxwell		34.66995	20	5m	

METEO FRANCE

Subject 3 Winner Slide 1

- **Données** : Relevés des stations météo de la veille + prédictions modèle physique
- **Traitement des données** : Sélection de tous les paramètres météo **moyennés** + **écart-types** associés (données station J-1 **+** données prévision)
- **Beaucoup de données NaN** : interpolation linéaire de ces données

METEO FRANCE

Subject 3 Winner Slide 2

Dataset d’entraînement

X_train

Données AROME

Complète les données manquantes

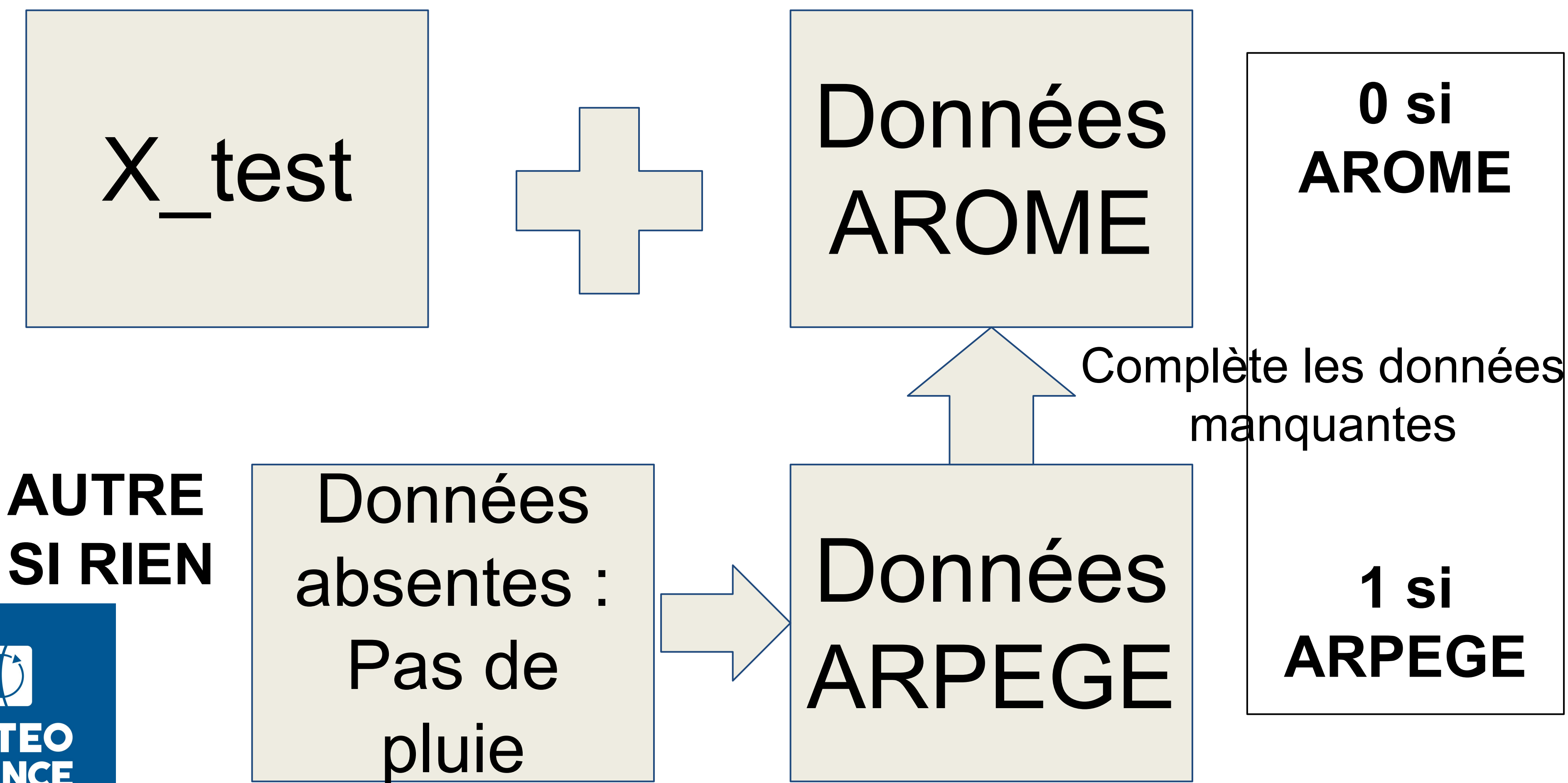
Données ARPEGE

0 si AROME

1 si ARPEGE

Subject 3 Winner Slide 3

Test



METEO FRANCE

Subject 3 Winner Slide 4

Model selection

- XGBoost avec MAPE en fonction objectif
- Optimisation des hyperparamètres par grid-search et cross-validation

```python
param_grid = {'param_1': [{'max_depth':5}, {'max_depth':6}, {'max_depth':7}], 'param_2':[25, 30, 35]}
```

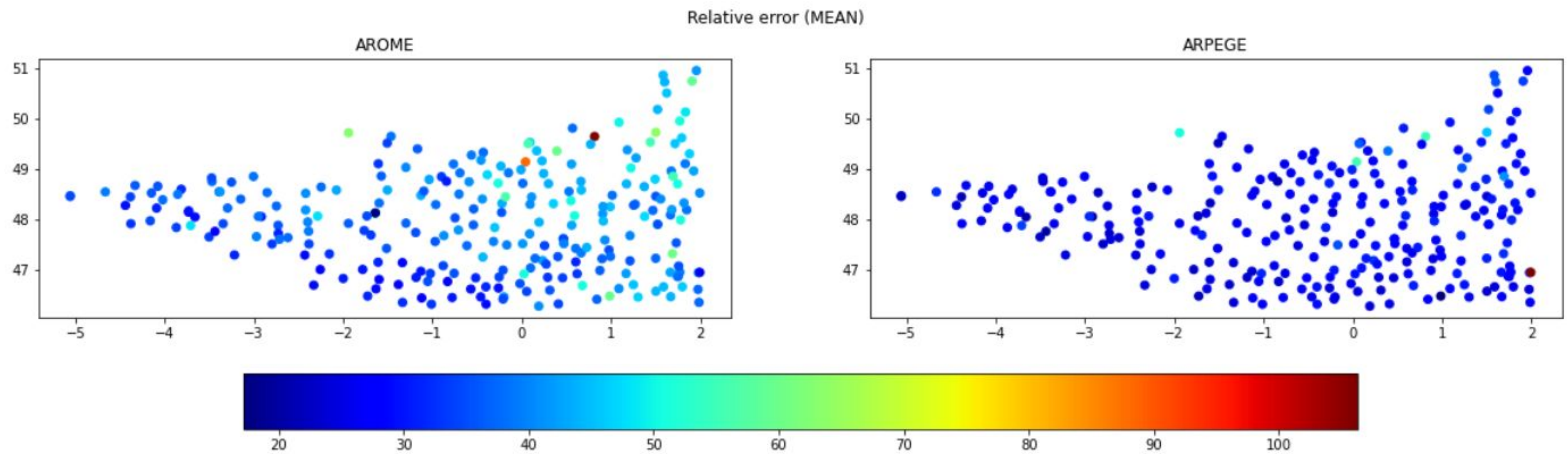
✓ 0.0s Python Python

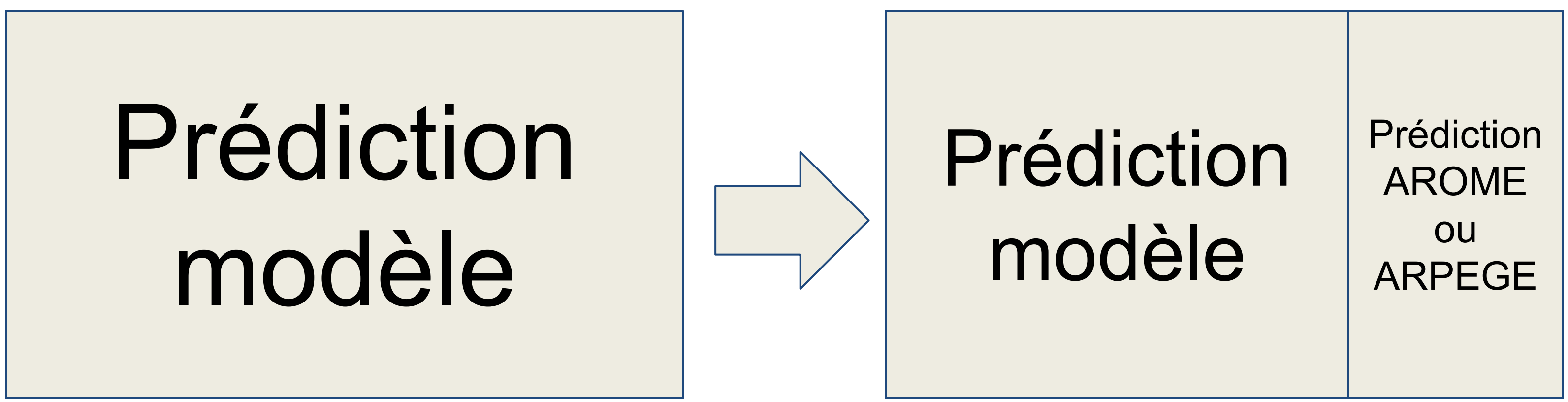
```python
# Compute cross-validation score
nb_trials = 4
score = []

for param_1 in param_grid['param_1']:
    for param_2 in param_grid['param_2']:
        mean_score = 0
        for i in range(nb_trials):
            Xtrain, Xval, ytrain, yval = train_test_split(Xtrain_full, ytrain_full, test_size=0.2)
            Xtrain_xgb = xgb.DMatrix(data=Xtrain, label=ytrain)
            Xval_xgb = xgb.DMatrix(data=Xval, label=yval)
            params = {'max_depth': 10}
            model = xgb.train(params=param_1, dtrain=Xtrain_xgb, obj=MAPE_obj, num_boost_round=param_2)
            ypred = model.predict(Xval_xgb)
            for i in range(len(ypred)):
                if ypred[i] < 0:
                    ypred[i] = 0
            print('*\n',end='')
            mean_score += MAPE(yval, ypred)
        print('max_depth:', param_1, 'n:', param_2, 'MAPE:',mean_score/nb_trials)
print(" done!")
```

✓ 49.7s Python

METEO FRANCE

Subject 3 Winner Slide 5





Subject 4: Maneuver detection

Responsable : Theo Nguyen, Dorian Gegout

Around **36500 Space objects** of more than 10 cm were referenced on December 2022. With the ramp-up of mega constellations, their number will increase exponentially. Any collision would threaten the viability of satellites at any orbit. It is critical for satellite operators to develop **collision avoidance algorithms**. Therefore, they need a precise satellite orbit determination. In this context, **Space Situational Awareness (SSA)** is an important space domain to get knowledge of space environment, including location and function of space objects.

Space debris artist view : source http://theglobserver.com)

Maneuver detection is part of SSA activities. Without knowledge of satellites maneuvers, classical algorithms used to determine satellites orbits could be imprecise or diverge. And here comes your contribution! Machine learning algorithms could provide an initial guess of the maneuver to refine orbit accuracy, and significantly improve the space traffic management.

Airbus Robotic Telescope (ART) is Airbus' own end to end capability for Space Surveillance & Tracking, performing automated optical observations of space objects from LEO to GEO.

From December 2021 to May 2022, the **Airbus Robotic Telescope** observed 3 satellites. From **irregular time series** of observations, your challenge will be to estimate the maneuver (detection, Δv, time of the maneuver …). You will design your algorithms on a simulated train dataset and confront them to the real test dataset.

AIRBUS DEFENCE & SPACE


Frosty Gould

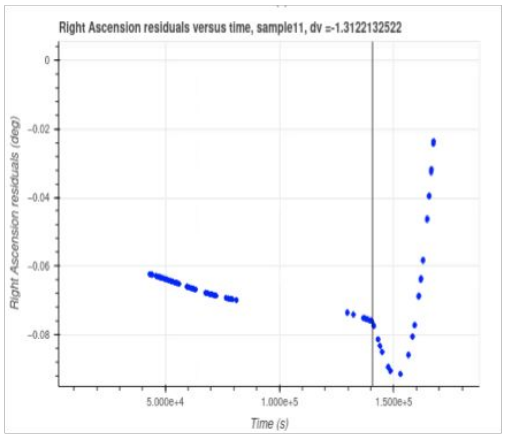

La PhotomentalTeam


X-PARA
2020
2021
See You in the Sky
Avada Kedavra

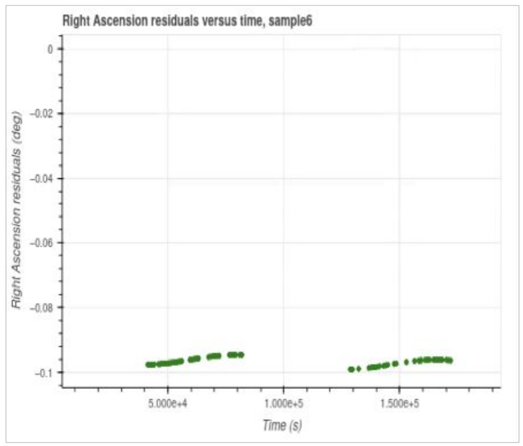
Problem

- After a maneuver, the residuals are impacted and a typical **pattern** is often visible.
- However, depending on the orbit determination errors, **residuals can be large** event without any maneuver. Therefore detection can be difficult.

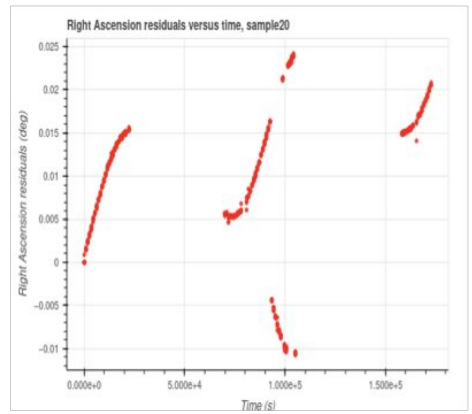
Airbus Robotic Telescope (ART)



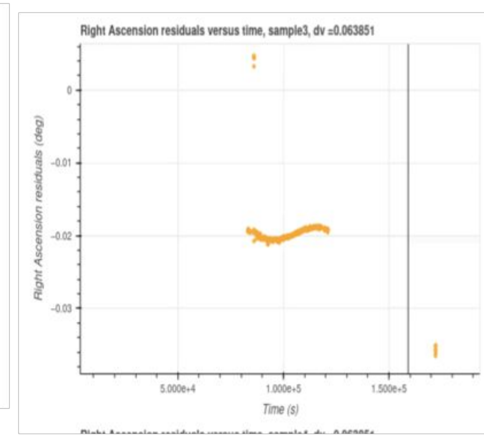
Right Ascension / Declination residuals with maneuver



Right Ascension / Declination residuals without maneuver



Residuals of false positive sample

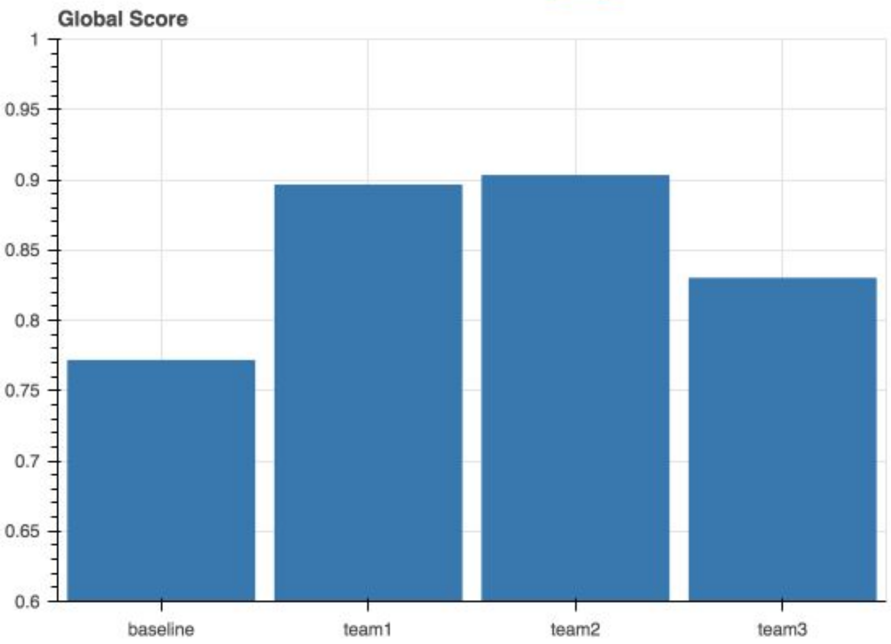


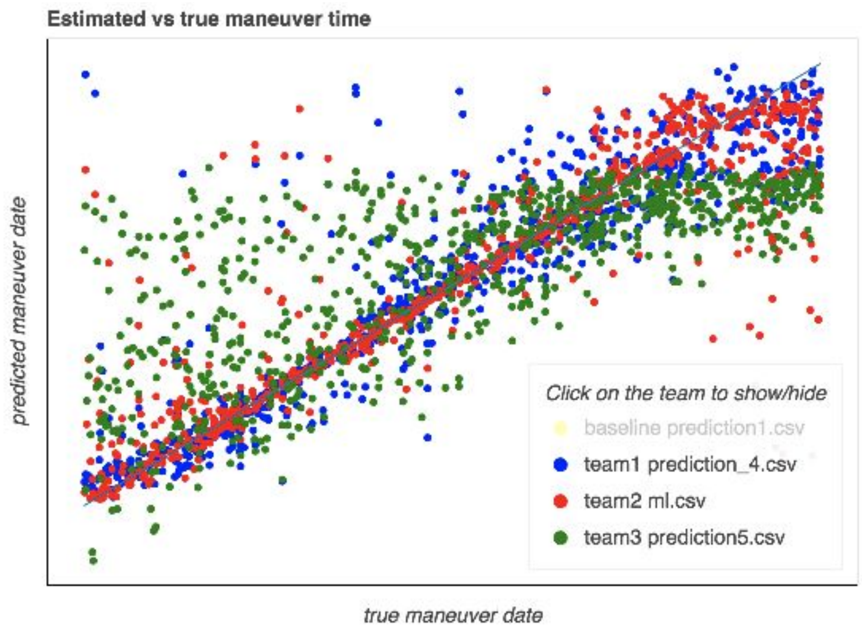
Residuals of false negative sample

AIRBUS DEFENCE & SPACE

Results

LEADERBOARD





First method : CNN

Three stages training approach :

1. Train simple {CNN + classification} head on classification task using the three features (ra, dec, sample time) ~ max F1-score: 0.81
2. Re-use CNN base net (e.g. remove classification head)
    a. { frozen CNN + regression head } to predict time of maneuver ~ MAE: 0.11
    b. { frozen CNN + regression head } to predict dV ~ MAE: 0.10
3. Unfreeze CNN and train full networks on regression task

Other approaches tried:

- CNN was also replaced with a temporal-CNN.
- Step 2 was sometimes useless for some specific data / architectures / hyper-parameters

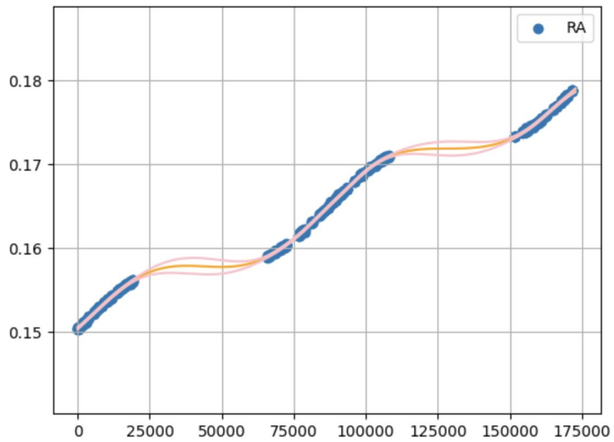
Preprocessing data :

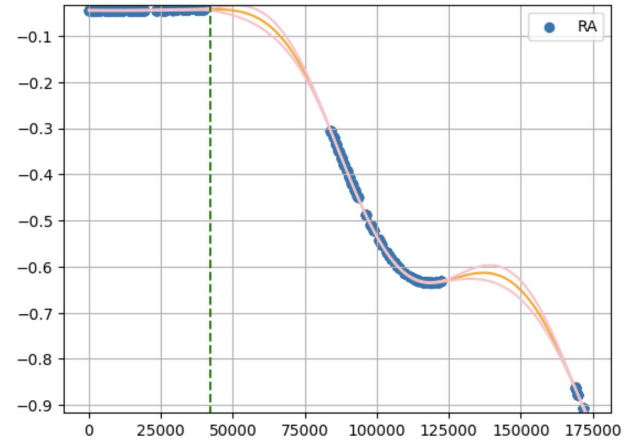
- Interpolation (presented afterwards). Was better on train set but overfitted too much.
- Remove times series for which the maneuver is at the very end and thus can not be predicted.
- Most likely other optim to investigate, for example using the delta time between each observation instead of the elapsed time since reference date. This would preserve shift-invariant.

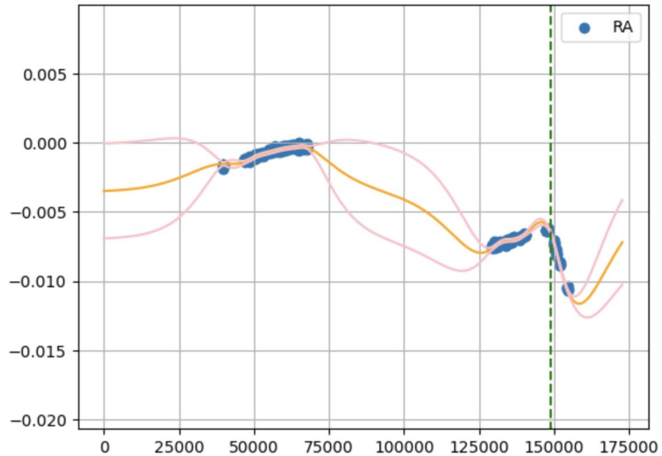
Conclusion: promising, especially with preprocessing, but requires fine-tuning of architectures and other-hyperparameters.

Interpolation

Gaussian Process Regression







AIRBUS DEFENCE & SPACE

Another Approach: Ensemble Methods

Classification

1 Feature (Right Ascension)
Interpolated Data

Method: Hard Voting Classifier
– Gradient Boosting Classifier
– Random Forest Classifier
– Gaussian Naive Bayes Classifier
– Extremely Randomized Trees Classifier

Regression (Maneuver Time)

1 Feature (Right Ascension)
Interpolated Data

Method: Stacking
– Gradient Boosting Regressor
– Random Forest Regressor
– Extremely Randomized Trees