

# DS P4 – Anticipez le retard de vol des avions



# Sommaire

Contexte  
Exploration  
Modélisation  
Conclusion

# Contexte - Objectif

## Modèle de prédiction des retard de vols d'avions

Evaluer les retards des vols américains 2016 à partir de variables fournies

### Les données

Dataset: 2016\_XX.csv (vols américain par mois 2016, 65 variables, 150Mo par fichier mensuel)

**Note:** l'étude porte sur les 3 premiers mois (1 350 000 observations) pour limiter les temps de calculs

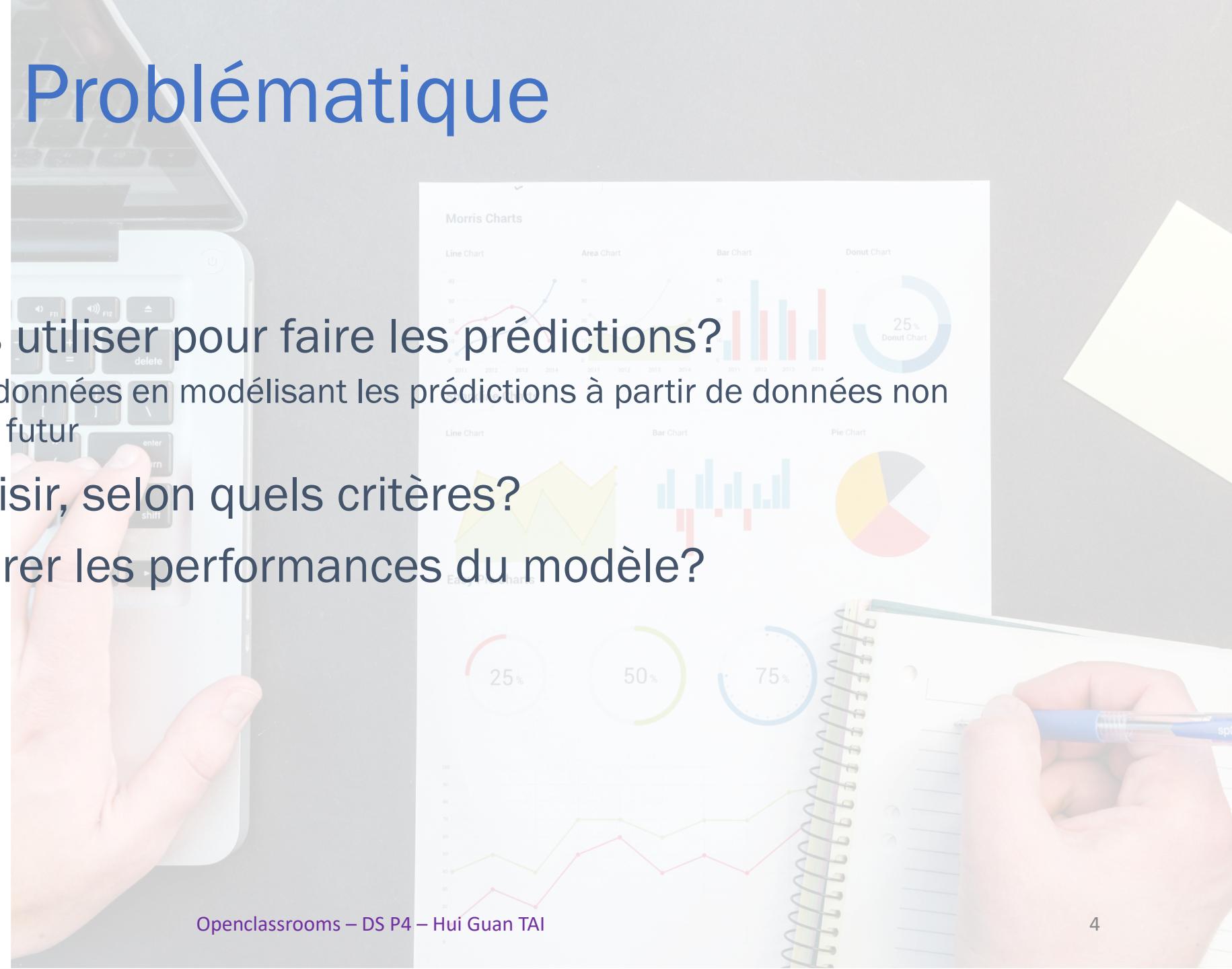
# Contexte - Problématique

Quelles variables utiliser pour faire les prédictions?

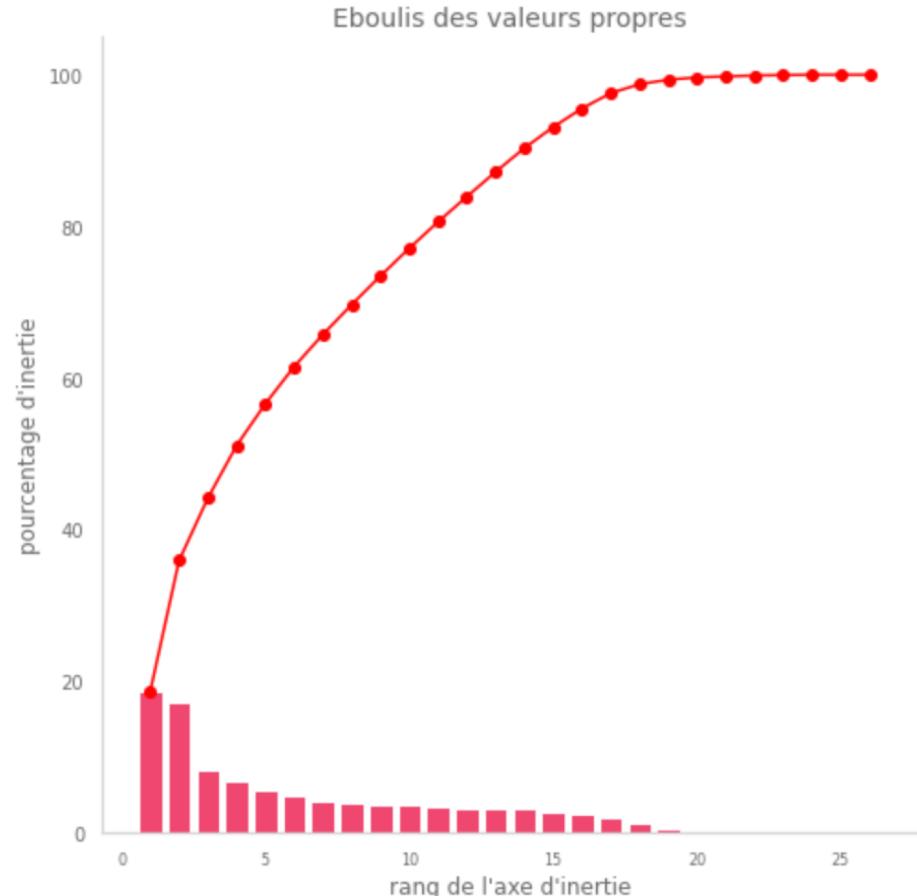
Eviter les fuites de données en modélisant les prédictions à partir de données non disponibles dans le futur

Quel modèle choisir, selon quels critères?

Comment améliorer les performances du modèle?



# Exploration – Familles de Variables



**La variable cible**

ARR\_DELAY : retard en minutes

**Familles de variables identifiées via ACP**

Temps écoulé (elapsed time)

Heure de départ (dep time)

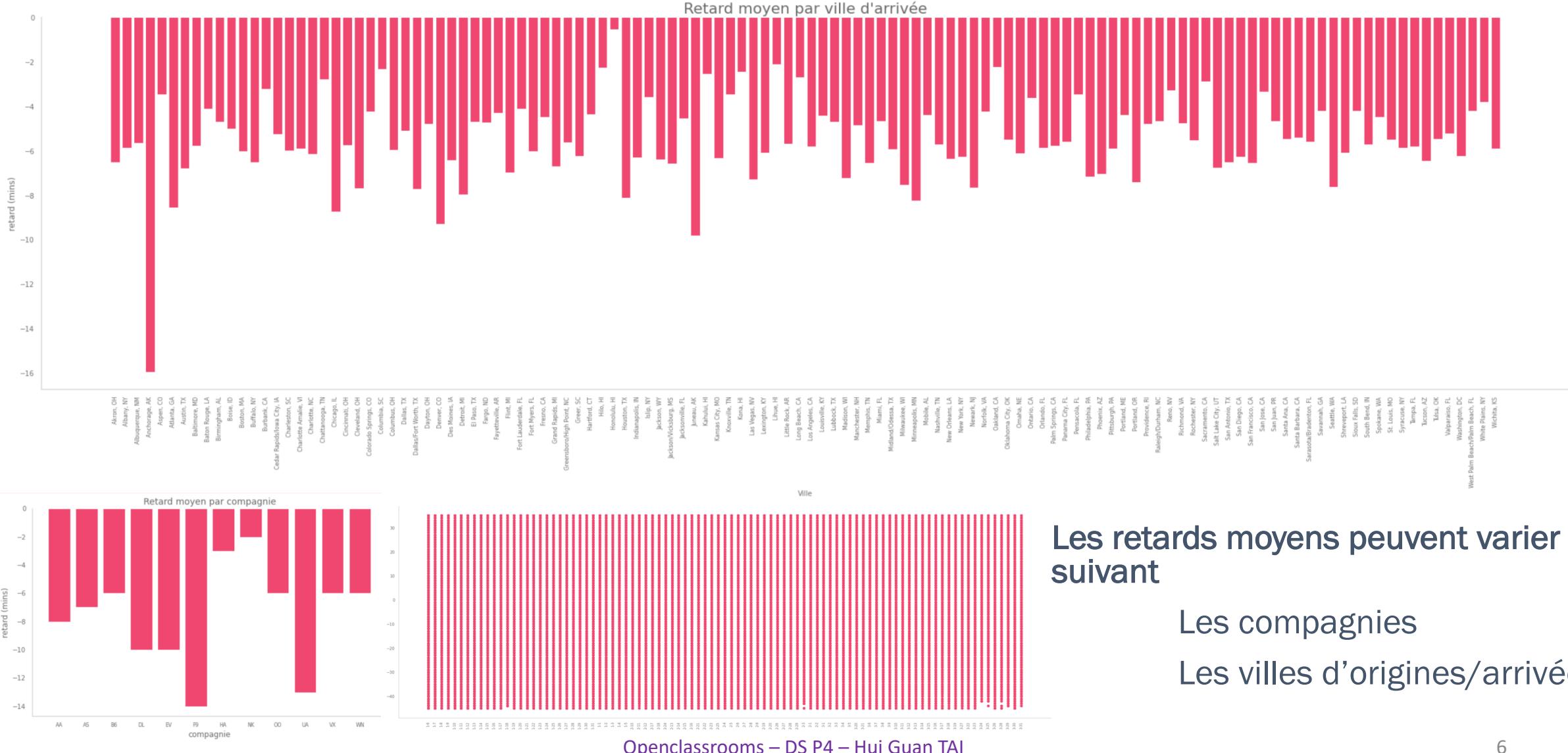
Retard au départ (delay)

Origine (origin)

Destination (dest)

Airline/Vol (airline/fl)

# Exploration – Moyenne des retards

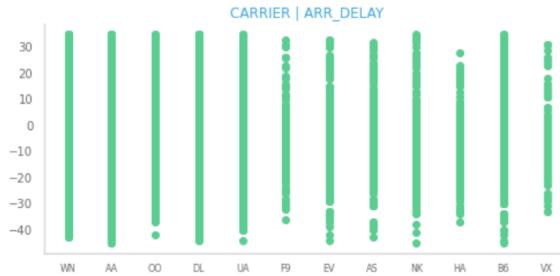
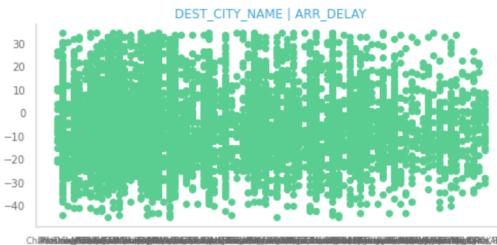
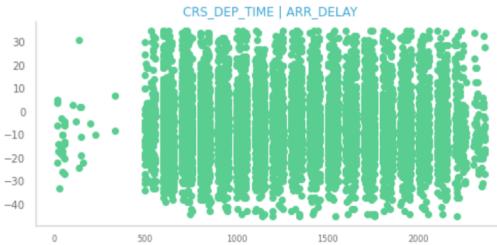


Les retards moyens peuvent varier suivant

## Les compagnies

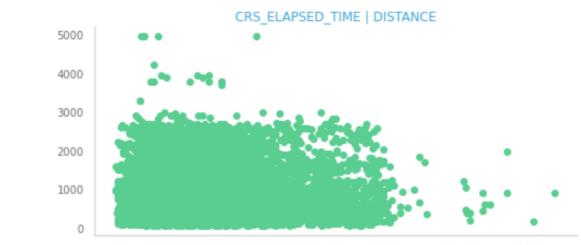
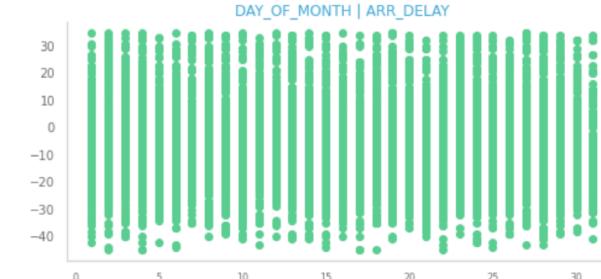
## Les villes d'origines/arrivées

# Exploration – Analyse bivariée



Distribution par paires entre le retard et  
La compagnie  
Le jour/mois  
L'heure de départ  
La ville d'origine/destination  
Durée de vol

=> Relation non linéaire



# Exploration – Encodage & Création de variables

1 - Get dummies sur les variables catégorielles CARRIER, ORIGIN\_CITY\_NAME, DEST\_CITY\_NAME

SHORT_FLIGHT	MEDIUM_FLIGHT	LONG_FLIGHT	DISTANCE	CARRIER_AS	CARRIER_B6	CARRIER_DL	...	DEST_CITY_NAME_San Juan, PR
0	1	0	601.00	0	0	0	...	0
0	1	0	1,618.00	0	0	0	...	0
0	1	0	2,106.00	0	0	0	...	0
0	1	0	1,616.00	1	0	0	...	0
0	1	0	1,028.00	0	1	0	...	0
0	1	0	447.00	0	0	0	...	0
0	1	0	1,005.00	0	0	0	...	0
0	1	0	2,282.00	0	0	0	...	0
0	1	0	874.00	0	0	1	...	0
0	1	0	628.00	0	0	0	...	0

=> total: 610 variables

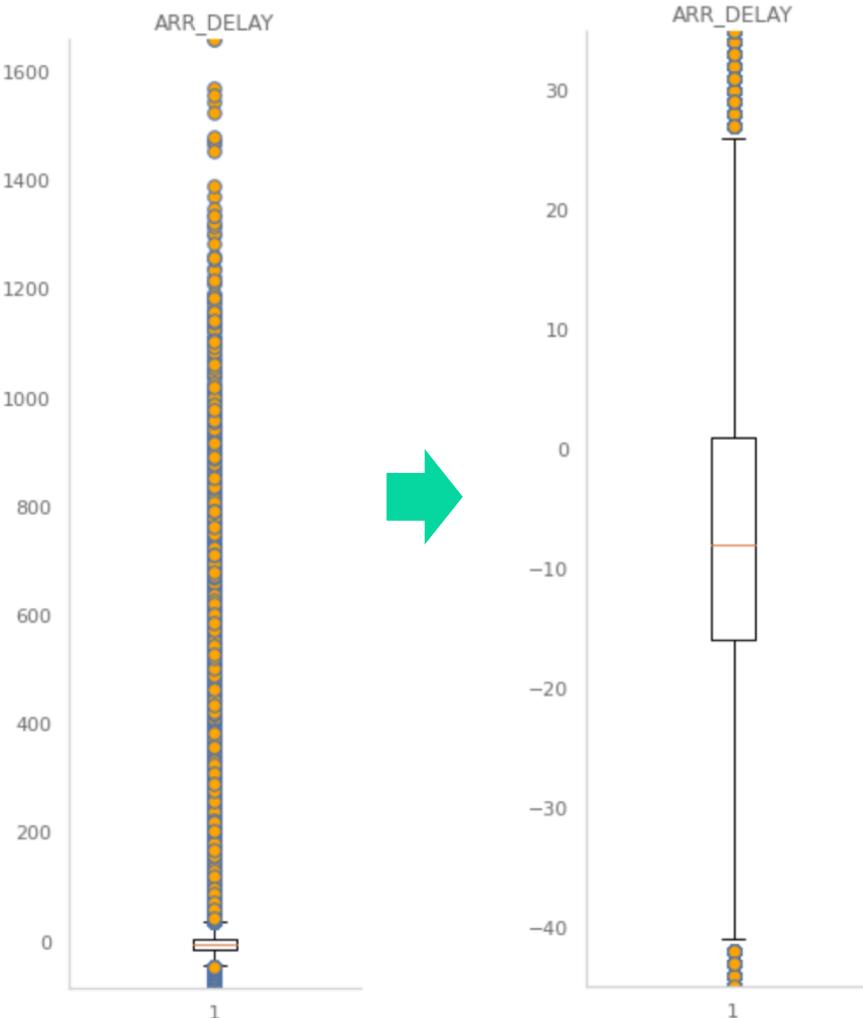
2 - On émet l'hypothèse que la distance peut influencer le retard

```
# Creation des variables Court, Moyen, Long Courrier
# <500km (311 miles), <5000km (3107 miles), <15000km (9321 miles)
dfCleaned['SHORT_FLIGHT']=(dfCleaned['DISTANCE']<310).astype(int)
dfCleaned['MEDIUM_FLIGHT']=((dfCleaned['DISTANCE']>=310) & (dfCleaned['DISTANCE']<3107)).astype(int)
dfCleaned['LONG_FLIGHT']=((dfCleaned['DISTANCE']>=3107) & (dfCleaned['DISTANCE']<9321)).astype(int)
```

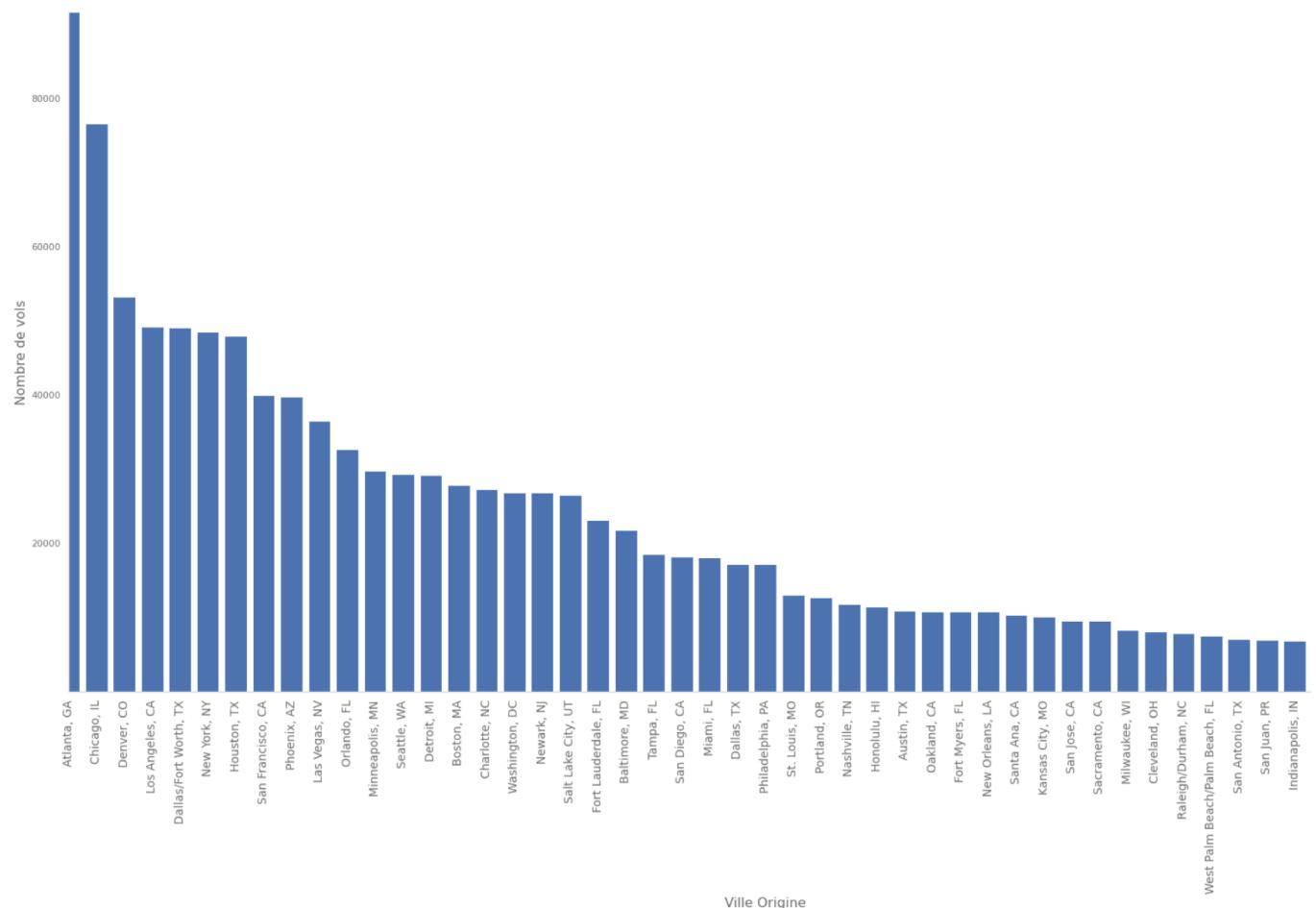
# Exploration – Suppression des atypiques

## 1 - Méthode des quartiles

(valeurs  $< Q1 - 1.5 \times IQR$  et  $> Q3 + 1.5 \times IQR$ )



## 2 - Filtrage des villes de destination et d'origine < 2000 vols



=> Total après suppressions : 170 variables

# Modélisation – Démarche

Séparation données d'entraînement/test

Choix de métriques

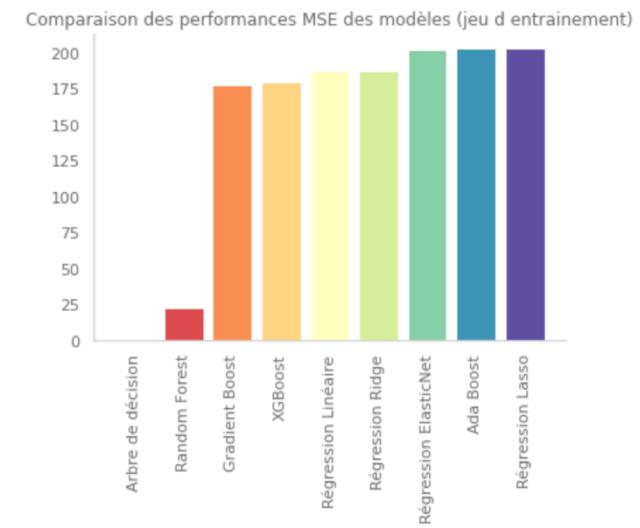
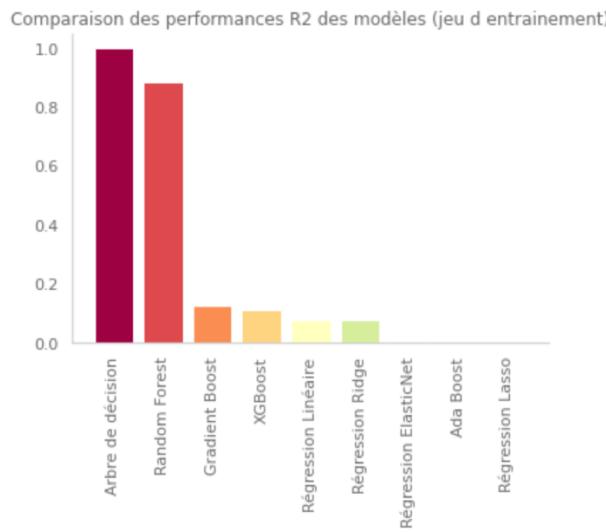
Comparaison des modèles

Optimisation & Validation croisée

Essai sur jeu de test

# Modélisation – Comparaison

Modèle	R2	MSE	Durée
Arbre de décision	1.00	0.01	28.18
Random Forest	0.62	77.40	603.80
Gradient Boost	0.13	177.63	431.68
XGBoost	0.12	179.87	94.61
Régression Linéaire	0.08	187.64	5.19
Régression Ridge	0.08	187.65	1.04
Régression ElasticNet	0.01	202.15	1.80
Régression Lasso	0.00	203.02	1.73
Ada Boost	0.00	203.23	204.24

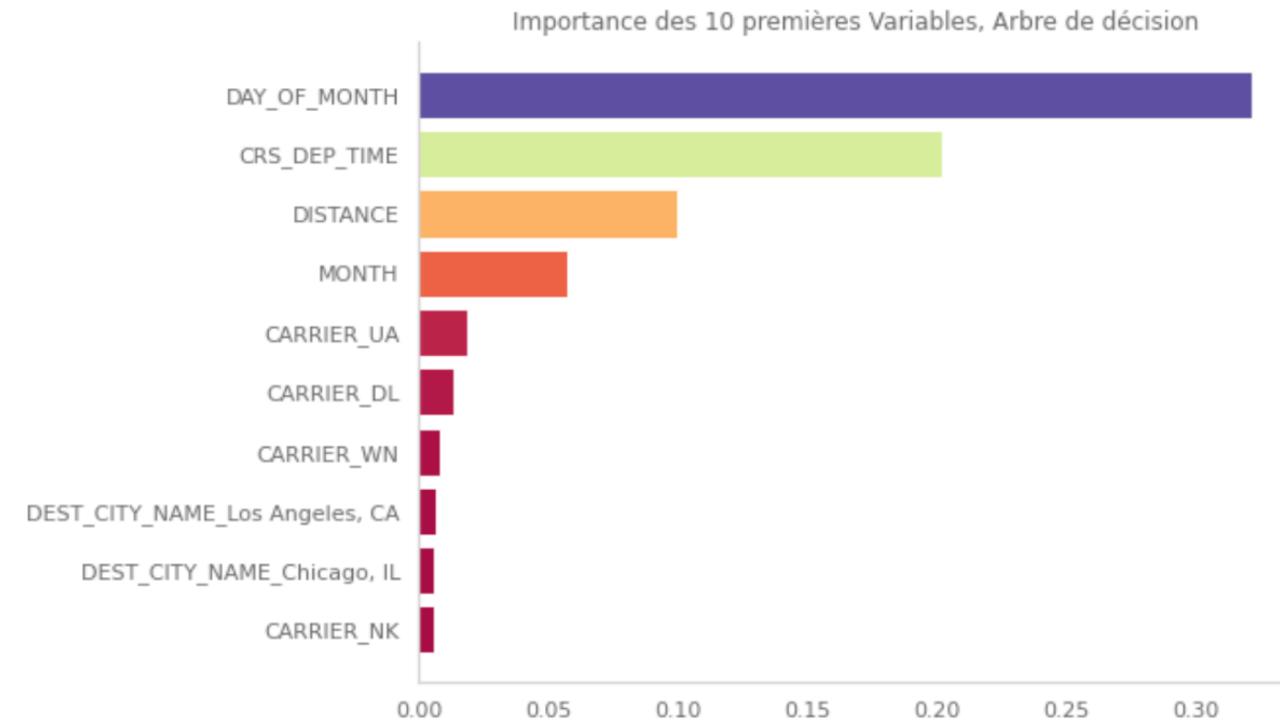


R2 pour mesurer le pourcentage d'explication de la variable par le modèle

MSE pour mesurer l'erreur de prédiction, sensible aux valeurs atypiques mais rapide (les valeurs atypiques sont déjà supprimées en amont)

=> le meilleur score R2, MSE pour l'arbre de décision et la forêt aléatoire

# Modélisation – Arbre de décision



## Optimisation Arbre de décision

Randomized search

Temps d'execution = 1 min 46 s

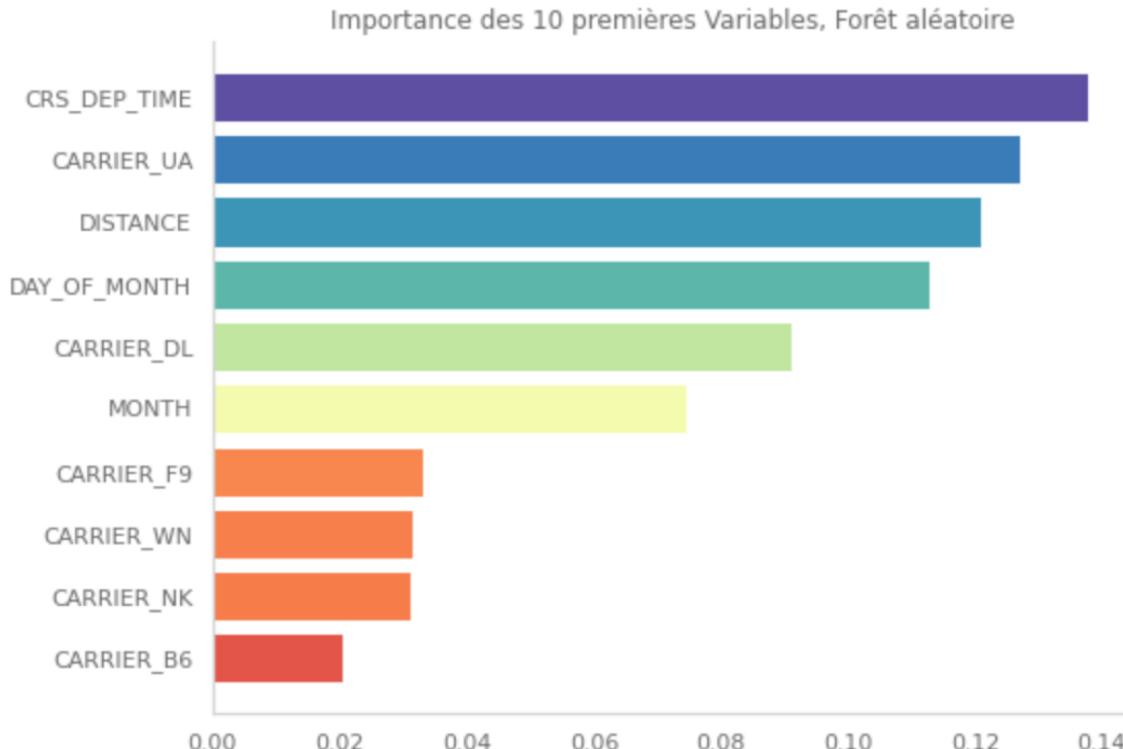
R2\_train= -0.21

MSE=-248

## Validation croisée 5 blocs

	0	1	2	3	4
fit_time	34.73	34.45	35.22	33.70	33.34
score_time	0.10	0.11	0.11	0.11	0.10
test_r2	-0.20	-0.20	-0.18	-0.18	-0.20
test_neg_mean_squared_error	-244.49	-243.45	-241.96	-239.07	-243.97

# Modélisation – Forêt Aléatoire



## Optimisation Forêt Aléatoire

Randomized search

Temps d'execution = 1 h 32 mins

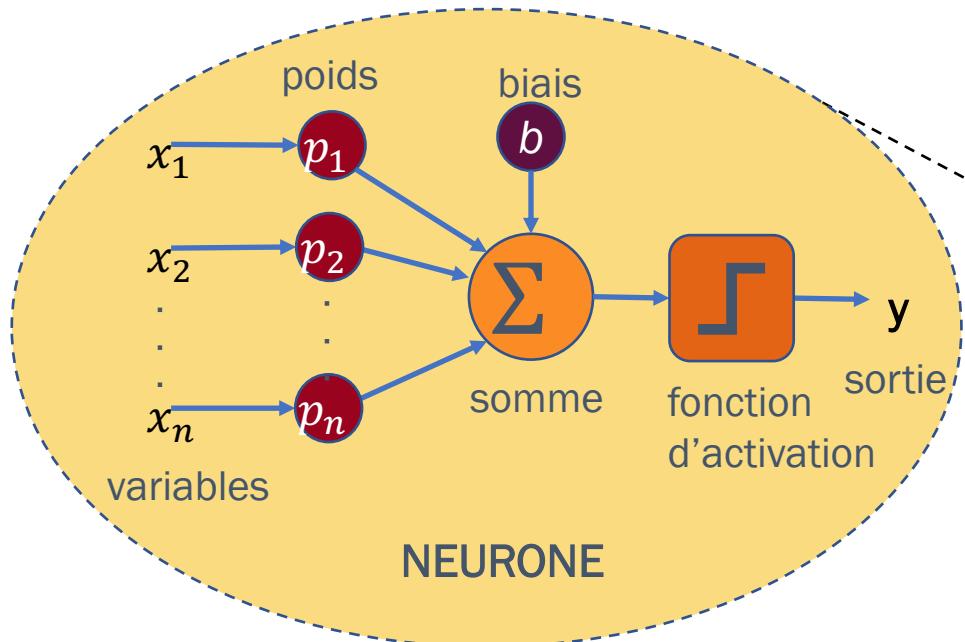
R2\_train= 0.18

MSE=-166.53

## Validation croisée 5 blocs

	0	1	2	3	4
fit_time	1,922.20	1,888.92	1,978.02	2,168.51	2,354.94
score_time	3.91	3.28	3.22	3.96	4.63
test_r2	0.19	0.19	0.19	0.19	0.19
test_neg_mean_squared_error	-164.90	-164.95	-165.66	-163.77	-164.51

# Modélisation – Réseau de neurones



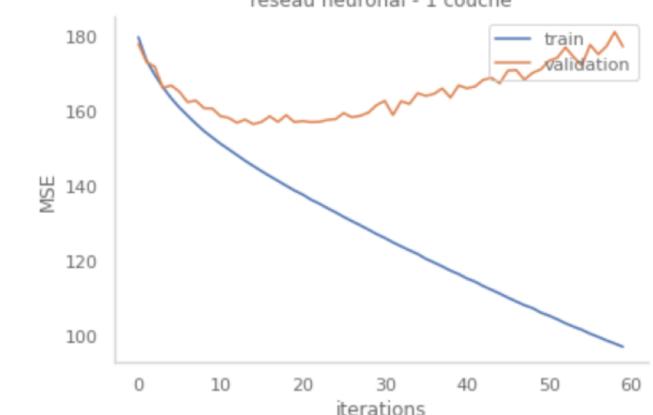
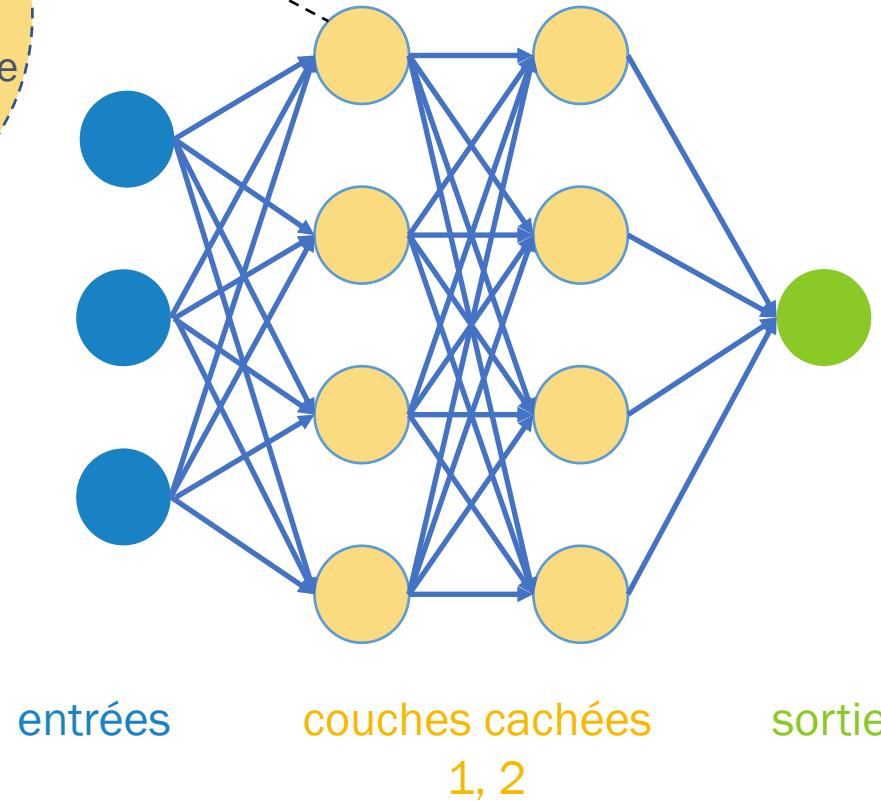
Couche entrée: 170 variables

Fonction d'activation (couches cachées): relu

Fonction d'activation (sortie): linear

Calcul de coût: MSE

Optimiseur: Descente de Gradient Stochastique

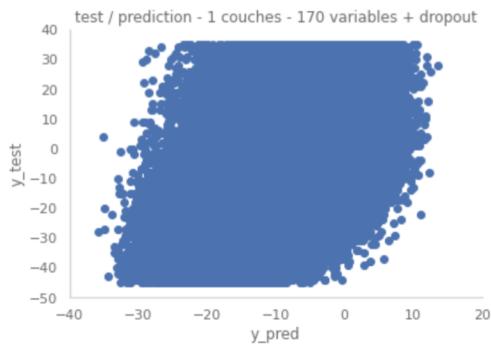


# Modélisation – Réseau de neurones

Nombre d'observations : 688 468

170 variables -> 2013 noeuds

## 1 couche

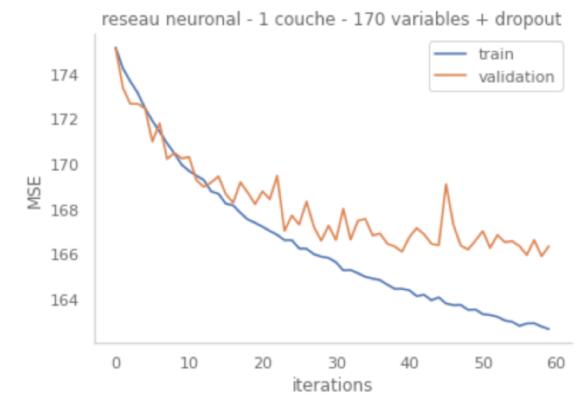


```
# 1 couche : 170 variables + drop out  
r2_score(y_train_tf,y_pred)
```

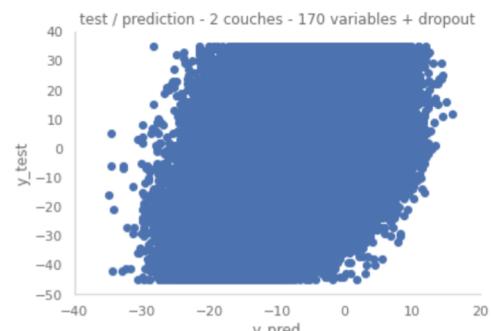
0.21304471525693003

```
# 1 couche sur les donnees test: 170 variables + drop out  
r2_score(y_test_tf,y_pred)
```

0.18160278933686969



## 2 couches

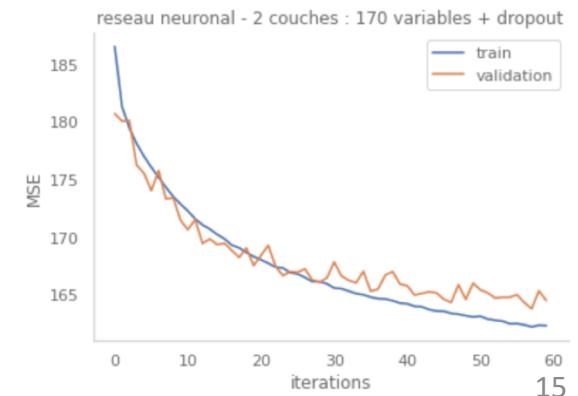


```
# 2 couches sur les donnees d'entrainement: 170 variables + dropout  
r2_score(y_train_tf,y_pred)
```

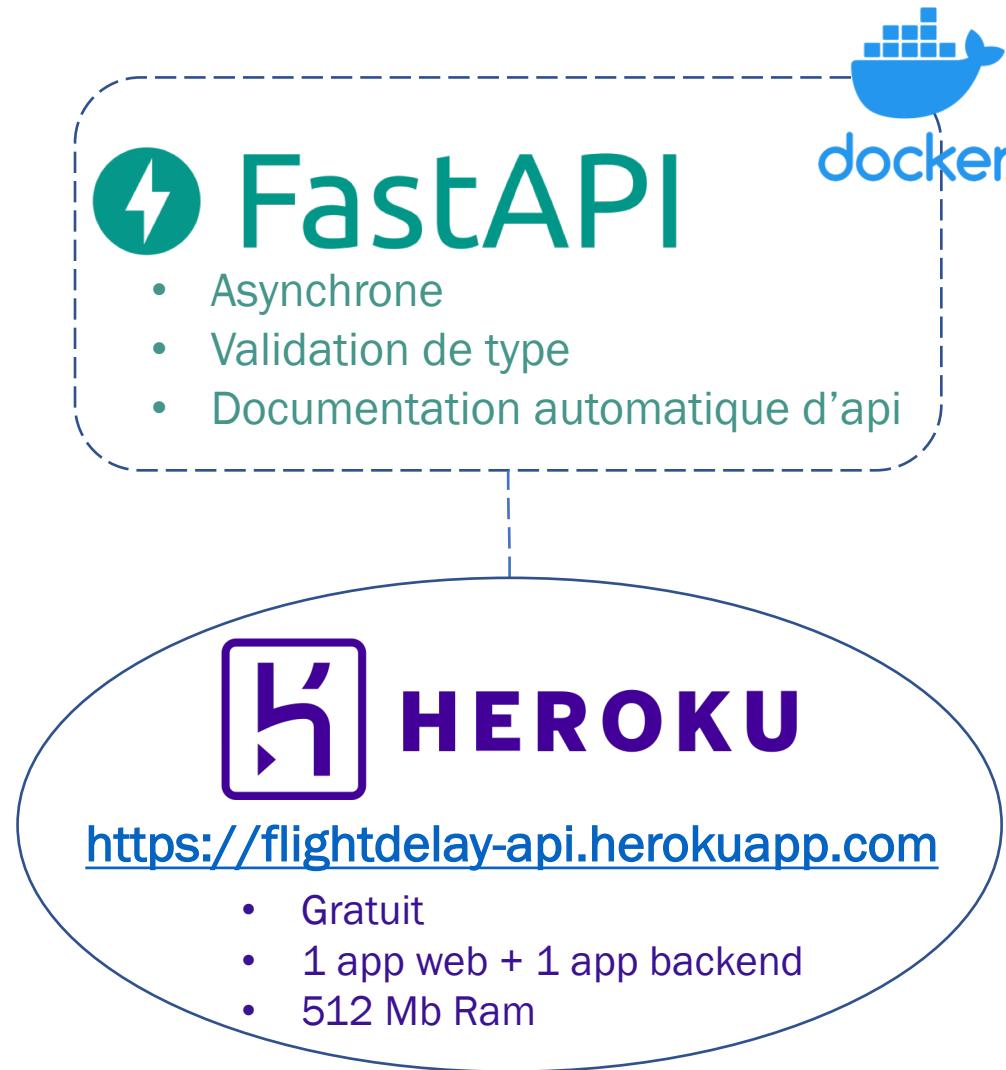
0.2260177203139695

```
# 2 couches sur les donnees de test: 170 variables + dropout  
r2_score(y_test_tf,y_pred)
```

0.1906550265265473



# Moteur - Déploiement



Anticipez les retards de vols US

The screenshot shows a user interface for predicting flight delays. The top section is titled "Paramètres de vol" (Flight Parameters) and includes fields for "Ville Départ" (Oklahoma City, OK), "Ville Arrivée" (Chicago, IL), "Compagnie" (EV), "Heure" (1757), "Jour" (6), and "Mois" (1). A "Submit" button is present. The middle section is titled "Données de test" (Test Data) and displays the query: "trajet de El Paso, TX à Dallas, TX avec WN le 1300/16/2". The bottom section is titled "Réponse" (Response) and shows a table of flight parameters:

664509	
ARR_DELAY	-5.0
CRS_DEP_TIME	1300.0
DAY_OF_MONTH	16
MONTH	2
DISTANCE	562.0
SHORT_FLIGHT	0
MEDIUM_FLIGHT	1
LONG_FLIGHT	0
CARRIER	WN
ORIGIN_CITY_NAME	El Paso, TX
DEST_CITY_NAME	Dallas, TX

The interface also includes a "Prédiction" (Prediction) section with a progress bar indicating "avance" (progress) at 5.762052 minutes.

# Conclusion

## Les -

- Le meilleur modèle de ce projet (réseau de neurones 2 couches) n'atteint qu'un coefficient de détermination  $R^2 \sim 0,2$

## Les +

- 1er modèles de prédiction
- Apprentissage et application de la démarche de modélisation
- Découverte et application de modèle Deep Learning

## Idées pour la suite

- Enrichir le dataset de données (autres mois, années, nouvelles variables)
- Essayer d'autres fonctions d'activations
- Transformer la problématique en sujet de classification (en retard/en avance)