28 oct 2025

1-) Construction d'un modèle : Sélection de prédicteurs. A56

2-) Dev Modèle dans un contexte Big Data : Spark ML. A57

3-) Déploiement de Modèle Concepts & Outils A61

TP : * Finalisation

* PPT

axis 1

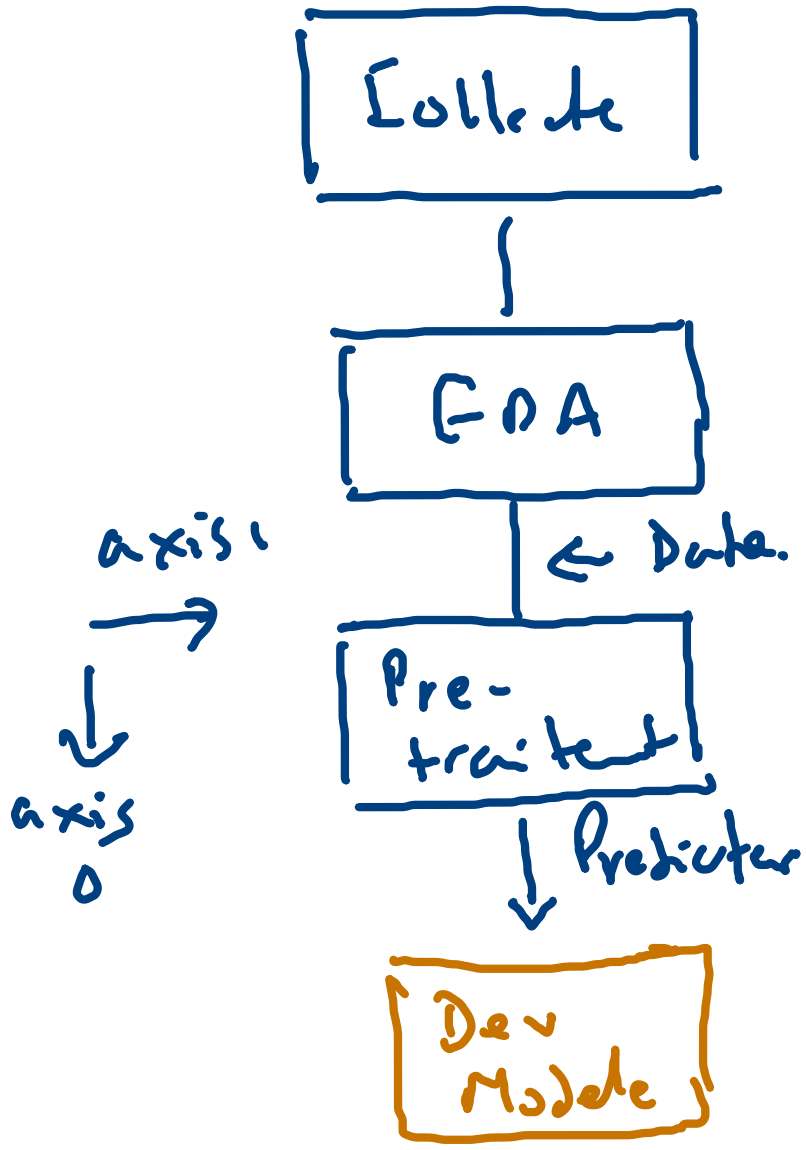
Descripteurs $\xrightarrow[\text{Selection}]{\text{Extraction}}$ Predicteurs

P.C.A : Reduction de Dimensions

$\sum p$ → p → Dev modele.

Tradeoff Biais-variance du modele

objectif : Prendre des predicteurs physiques
⇒ offrir un nombre plus reduit de predicteurs



All-in ~~IS~~ ~~non~~ poids Taille N Ecole QI → All-in → y (note)

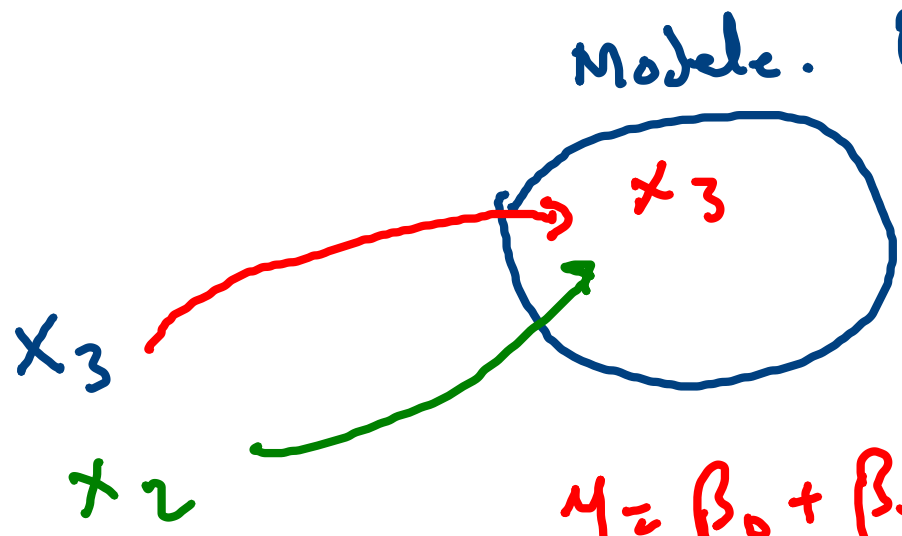
? Question: Poids, Taille n'ont pas d'effet sur le y (note)

- Corrélation.
- Expertise du domaine

Performance du modèle : À voir.

Stepwise Forward

X_1 X_2 X_3 X_4



$y = \beta_0 + \beta_1 X_3$

(?) Qualité < BL

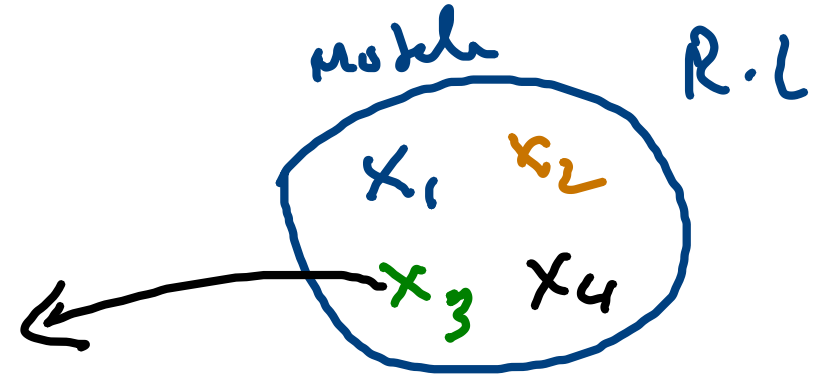
$y = \beta_0 + \beta_1 X_3 + \beta_2 X_2$

Qualité < BL

>

stepwise Backward.

X_1 X_2 X_3 X_4



$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Qualité < BL

$y = \beta_0 + \beta X_1 + \beta_2 X_2 + \beta_4 X_4$

Qualité > BL

Backward it1 x_1 x_2 x_3 x_4 $\alpha = 0.05$

p-value. 0.01 0.015 0.07 0.03

it2 Model → x_1 x_2 x_4

p-value 0.013 0.053 0.02

Model x_1 x_4 ⇒ Model

0.027 0.032 ↳ x_1 x_4.

E-D-A

Dist

x_1 | x_2 | ... x_n | y

Standard Scale
Robust Scale.
MinMax Scale.

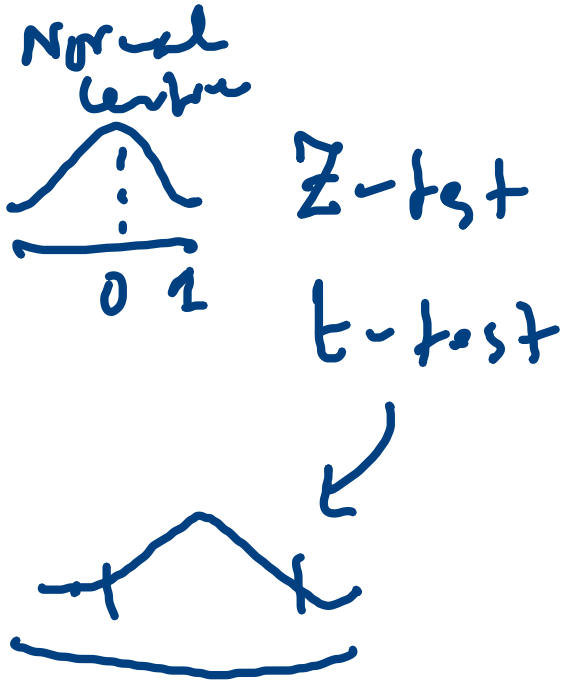
Test d'Hypothèse.

H_0 : X_i n'a pas d'effet sur y — Absence.

H_1 : X_i a un effet sur y — Non Absence.

Calculer une statistique → Dépend de la dist X_i

$\alpha = 0.05$



p-value : Mesure de prob pour indiquer si on est dans les conditions de H_1,

p-value $|calculée| < 0.05 (\alpha)$

$\Rightarrow H_1$

Forward.

iteration 1.

x_1 —— Mod1 p-val 0.03

x_2 —— Mod2 p-val 0.04

x_3 —— Mod3 p-val 0.07

x_4 —— Mod4 p-val. 0.01

Min

Mod 1T1

x_4

x_4 x_1 0.02

x_4 x_3 0.06 ✗

✓ x_4 x_2 ?! 0.01.

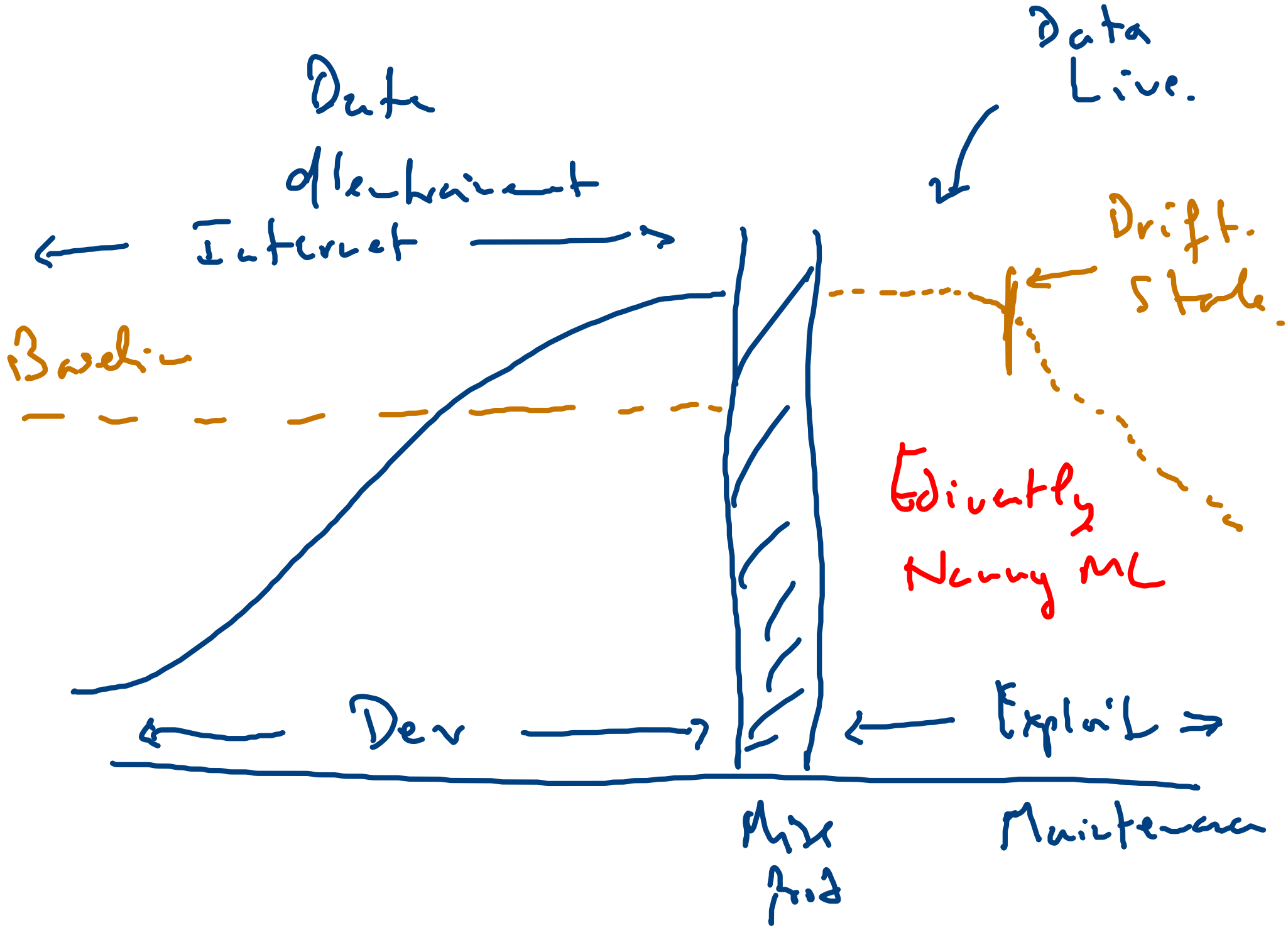
Mod 1T2

x_4 x_2

FIN

✗ x_4 x_2 x_1 0.055 ≥α

x_4 x_3 x_2 0.07 ≥α


Data
Live.
Data
Deployment
Internet
Drift.
Baseline
Evidently
Nanny ML
Dev
Exploit
Mix
Maintenance

	x_1	x_2	x_N	y
test	—	—	—	—

train-test

y_{test} connu.

y_{pred} → à partir du modèle

Calcul ← Met(y_{pred}, y_{test})

Prod

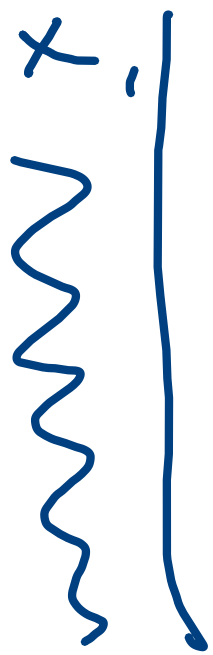
	x_1	x_2	x_N	y_{pred}
	—	—	—	

y_{pred} connu

y_{real} ??

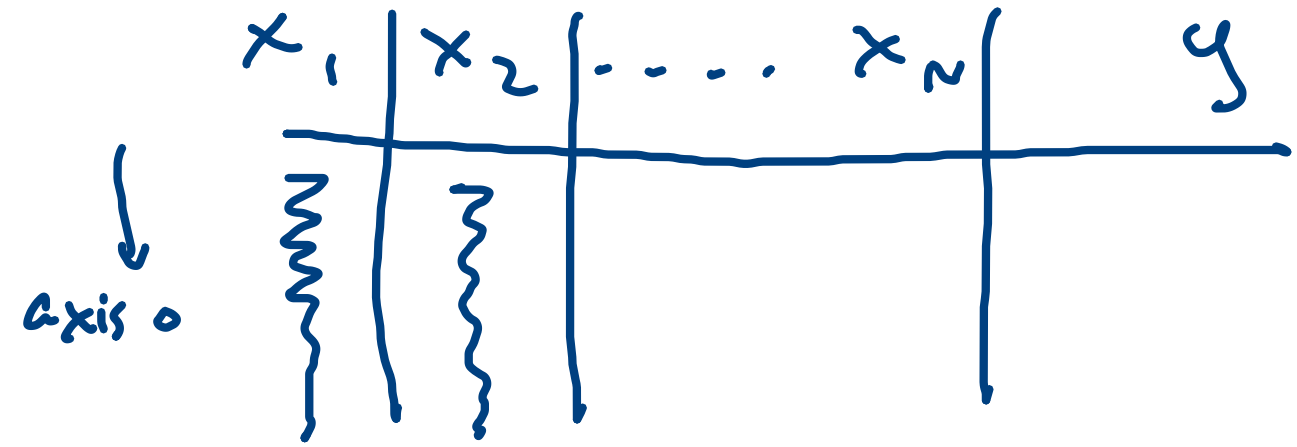
Approach Calcule Metrique

train



D.P train ⟵ Outil ⟶ D.P prod

Process



E.D.A 1) Analyse statistique Descriptive — moy, median. ecart type, etendue min, max

3) Apporte des corrections.

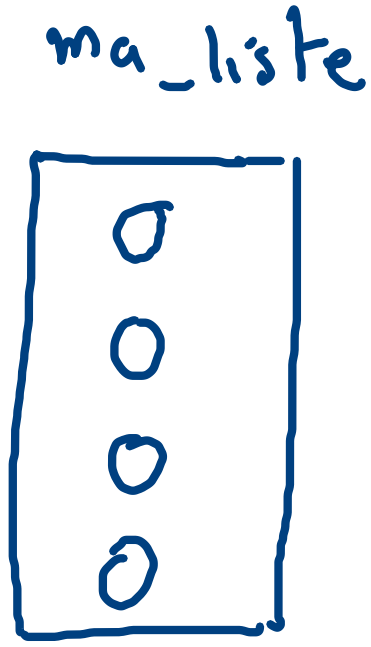
2) Dist Prob $\{x_i\}$

4) Tests d'hypothèse — H_0 : x_i pas d'effet sur y
H_1 : x_i a effet sur y

$$

Statistique : Test paramétrique — Z-test, t-test ✓
Test non paramétrique

Spark

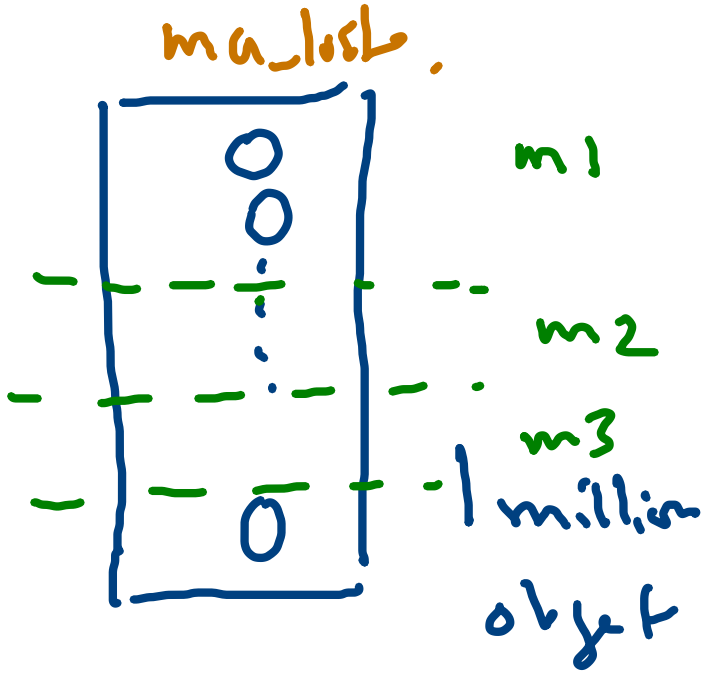



```
for tmp in ma_liste :
    tmp.nom = tmp.nom.upper()
```

} T_M

non solaire.

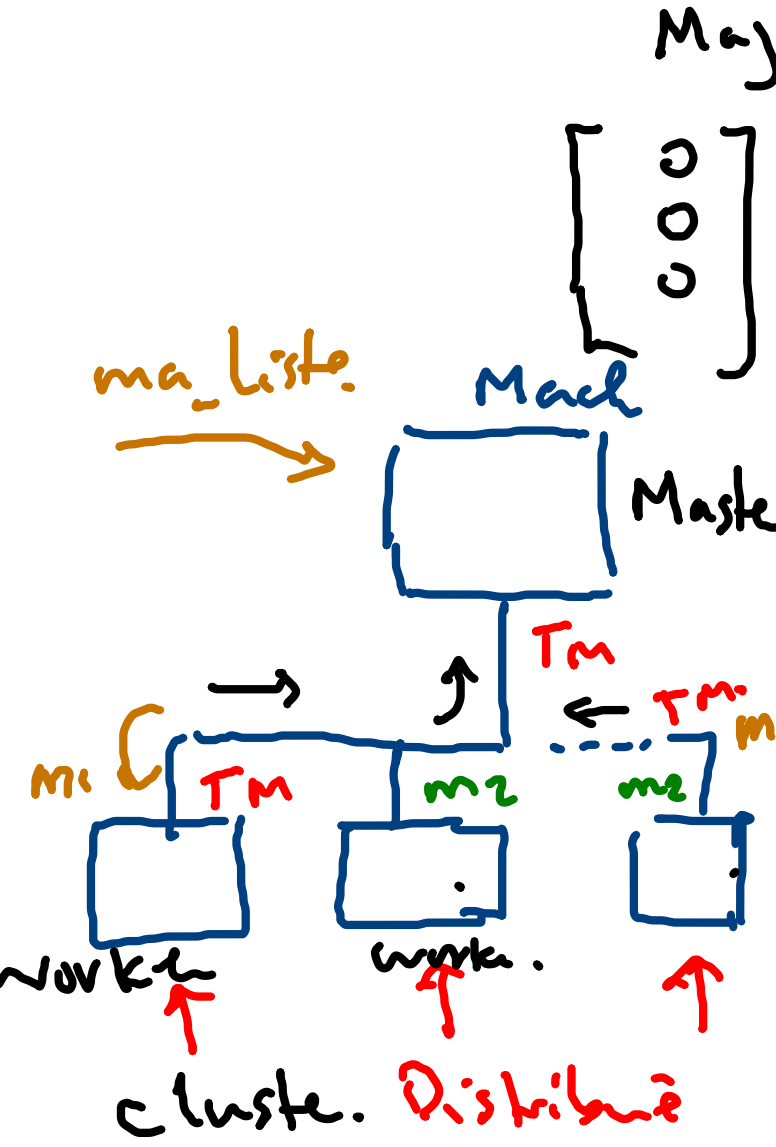
Durée traitement



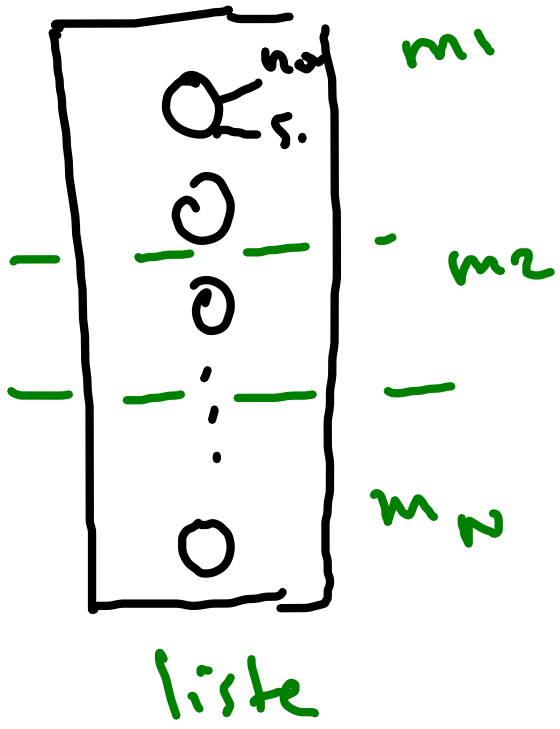
T_M [0 0 ... 0]

T_M [0 ... 0]

T_M [0 ... 0]

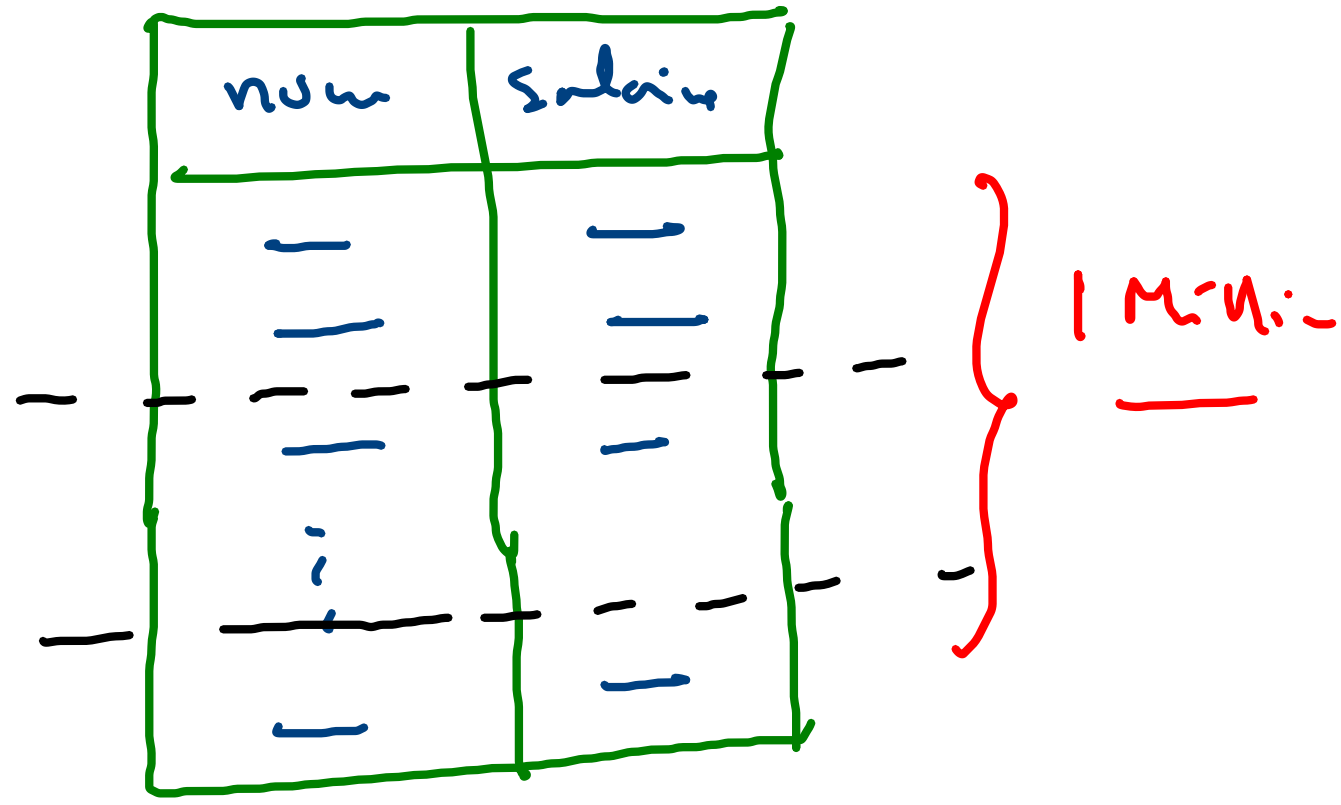


Framework Spark



liste

RDD



DataFrame. Spark

↳ Représentation ML

SparkML

Scikit-learn

↙ RL

KNN

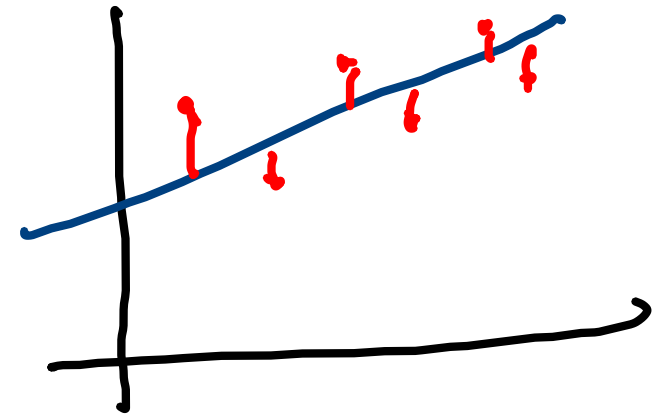
⋮

Datatron.

RL

fc loss

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_p^i - y_r^i \right)^2$$



$$\Rightarrow \hat{\beta}_i = \frac{1}{N} \sum_{i=1}^{1000} (\quad) + \sum_{i:1001} \quad + \cdots\cdots$$