

Formation

Machine Learning avec Python

08/11/2023



VOTRE **FORMATEUR**

Cédric DANGEARD

Consultant Data

cedric.dangeard@businessdecision.com

SOMMAIRE

- 8 novembre 2023
 - Introduction
 - Le ML en entreprise
 - Projet Data Science
 - Etapes d'un projet de Machine Learning
 - TD : Préparation des données
- 22 novembre 2023
 - Bagging & Boosting
 - TD : Mise en place d'un modèle
- 29 novembre 2023
 - Optimisation
 - TD : Optimisation
 - Projet final



#Data #DataDriven
#DataStrategy #DataThinking
#DataTraining
#ChangeManagement
#SelfServiceData #BigData
#DataArchitecture #Cloud
#DataIntegration #ETL #BI
#DataEngineering
#ArtificialIntelligence
#AdvancedAnalytics
#DataScience #MachineLearning
#FraudAnalytics
#DataIntelligence
#ReportingDashboarding
#DataGovernance #DQM
#MDM #DataCatalog #EPM
#Digital #CRM #IoT #Chatbot
#Website
#MarketingAutomation
#eCommerce #MobileApps
#DataSecurity #IAM

INTRODUCTION

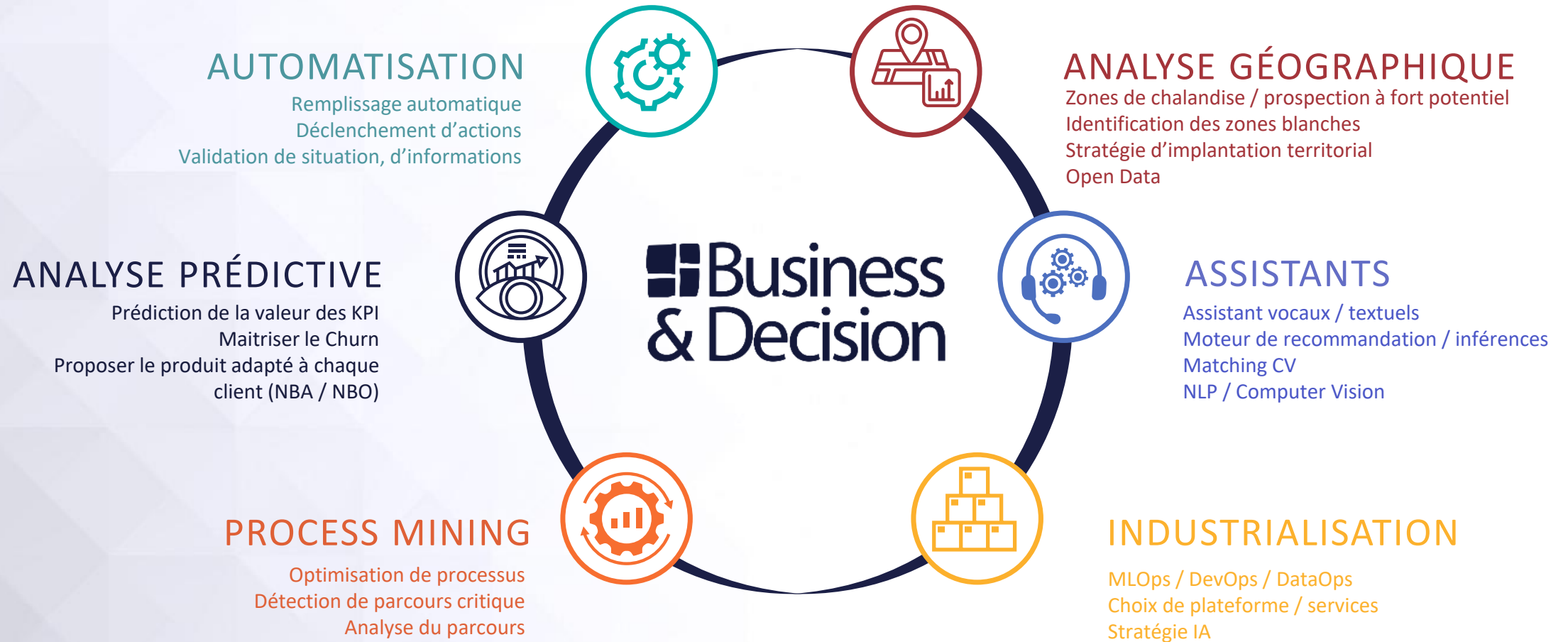
Science des données (Wikipédia) : Étude de l'extraction automatisée de connaissance à partir de grands ensembles de données.

Leur apparition repose sur plusieurs faits :

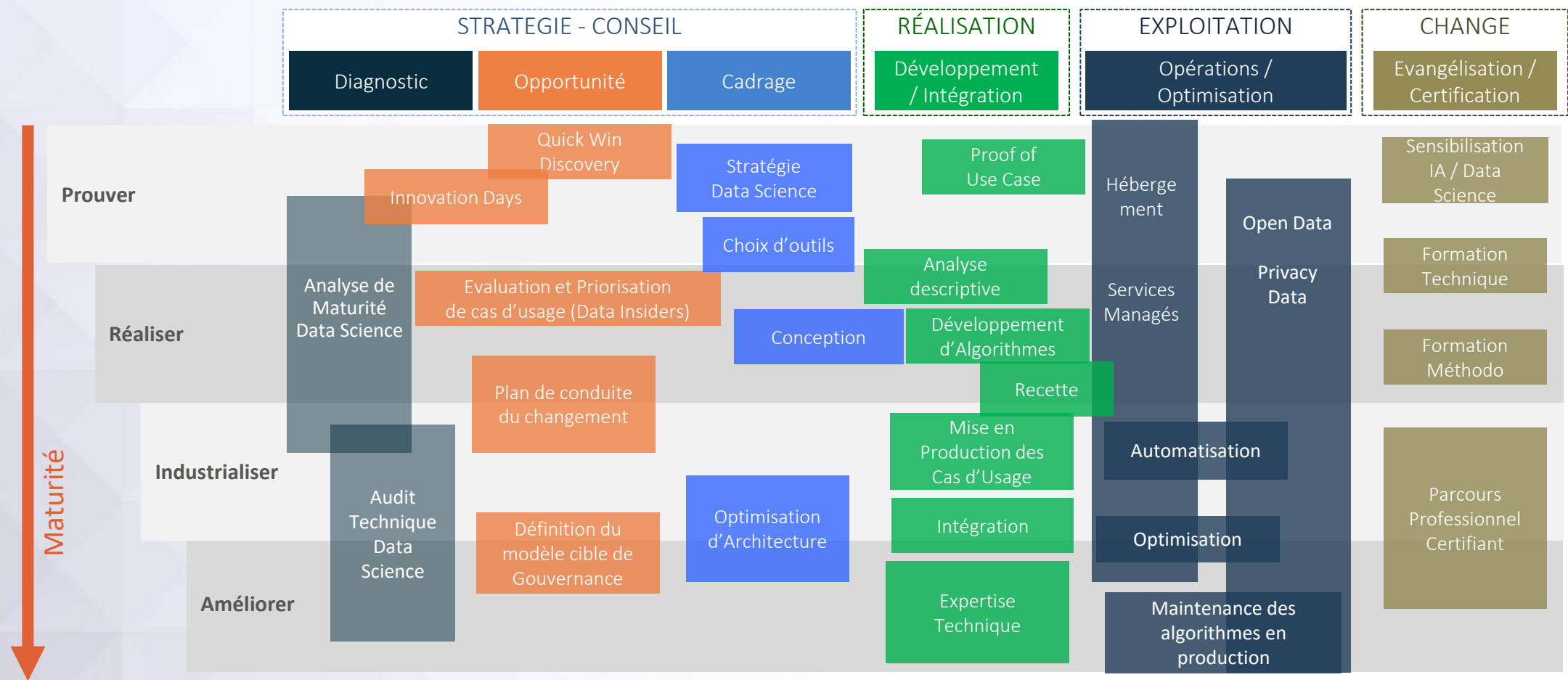
- L'explosion du volume de données et de la puissance machine
 - La diversification des types de données (image , textes...)
 - Le développement de l'Open source
- Aujourd'hui encore la plupart des données interprétées par une entreprise reste de la donnée structurée ...
 - ... Alors que plus de **80%** des données générées sont non structurées, ce qui représente une perte d'informations précieuses.

ML EN ENTREPRISE **MACHINE LEARNING**

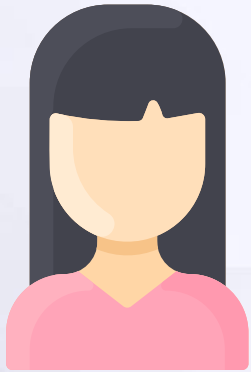
DOMAINES DE L'IA QUE L'ON COUVRE



CATALOGUE DE SERVICES DATA SCIENCE & IA



QUELLES DIFFÉRENCES ENTRE LES PROFILS ?



Data Analyst

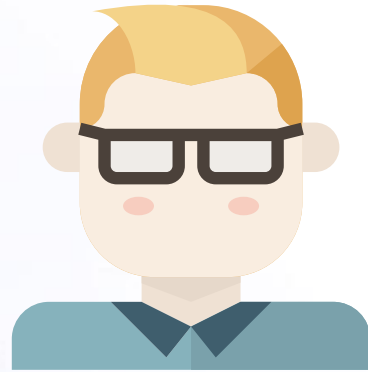
Technique

Métier

Statistique

Optimisation

Analytique



Data Engineer

Technique

Métier

Statistique

Optimisation

Analytique



Data Scientist

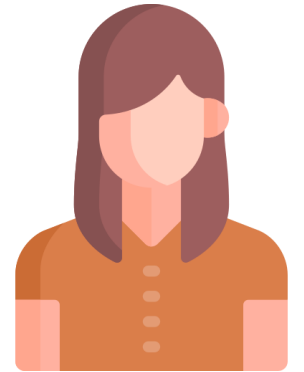
Technique

Métier

Statistique

Optimisation

Analytique



ML Engineer

Technique

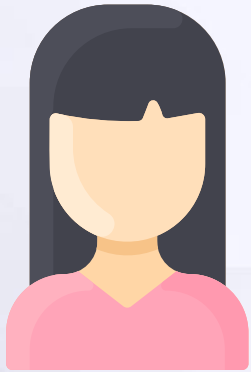
Métier

Statistique

Optimisation

Analytique

QUELLES DIFFÉRENCES ENTRE LES PROFILS ?



Data Analyst

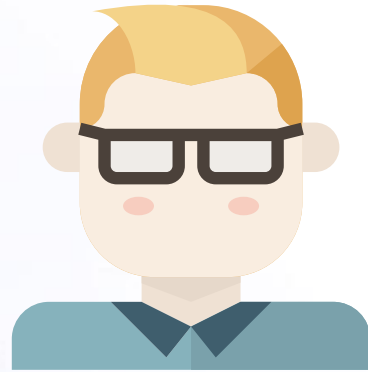
Technique ★★

Métier ★★★★★

Statistique ★★

Optimisation ★

Analytique ★★★★★



Data Engineer

Technique ★★★★★

Métier ★

Statistique ★★

Optimisation ★★★★★

Analytique ★★



Data Scientist

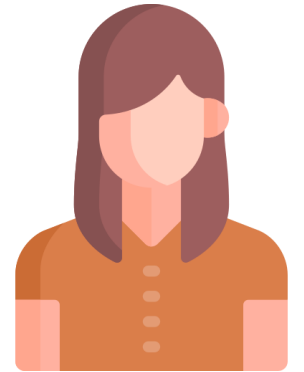
Technique ★

Métier ★★★★★

Statistique ★★★★★

Optimisation ★★★

Analytique ★★



ML Engineer

Technique ★★★★★

Métier ★

Statistique ★★★★★

Optimisation ★★

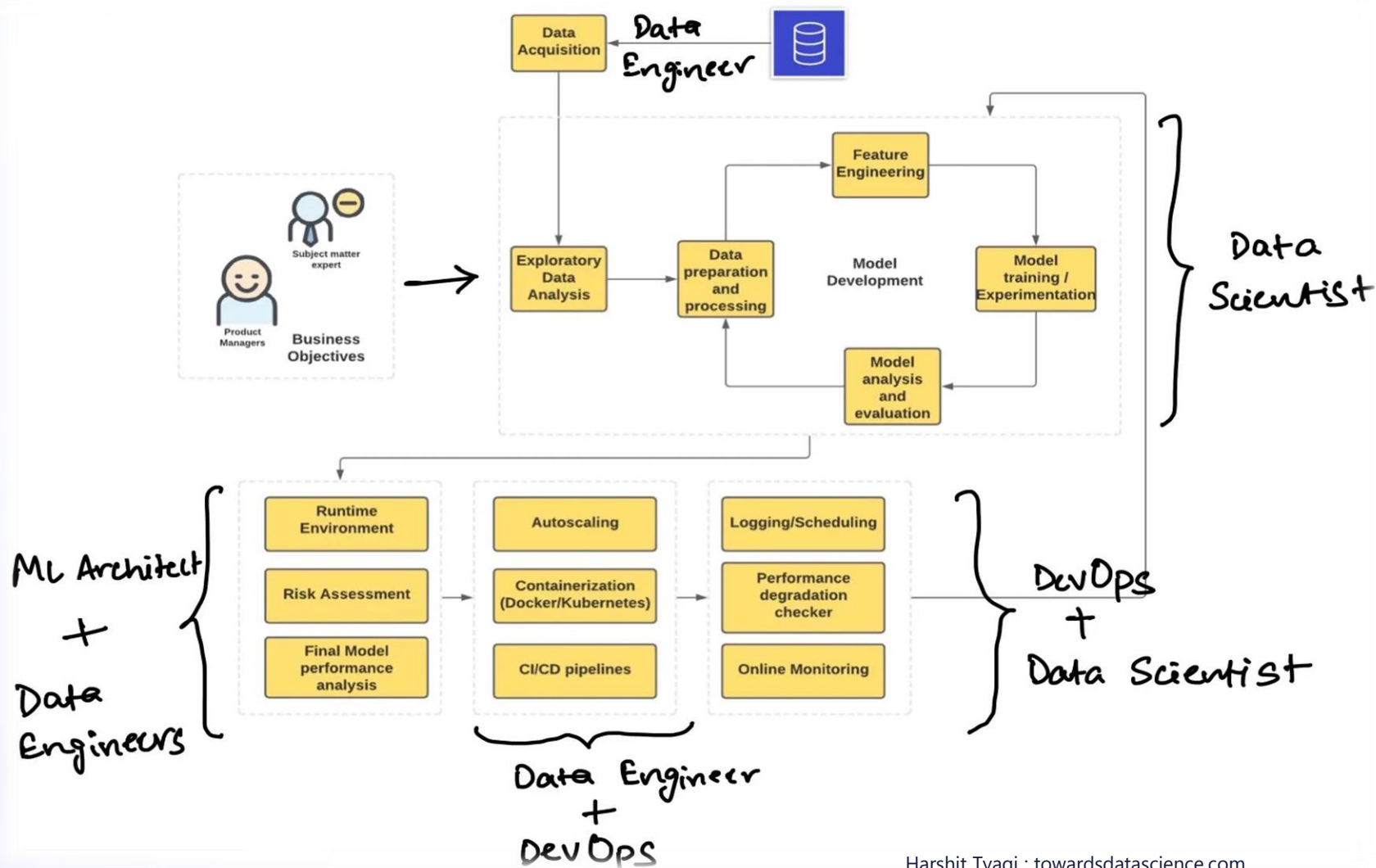
Analytique ★

ML OPS

Un Cycle de Vie

- Développement
- Déploiement
- Monitoring
- Retraining

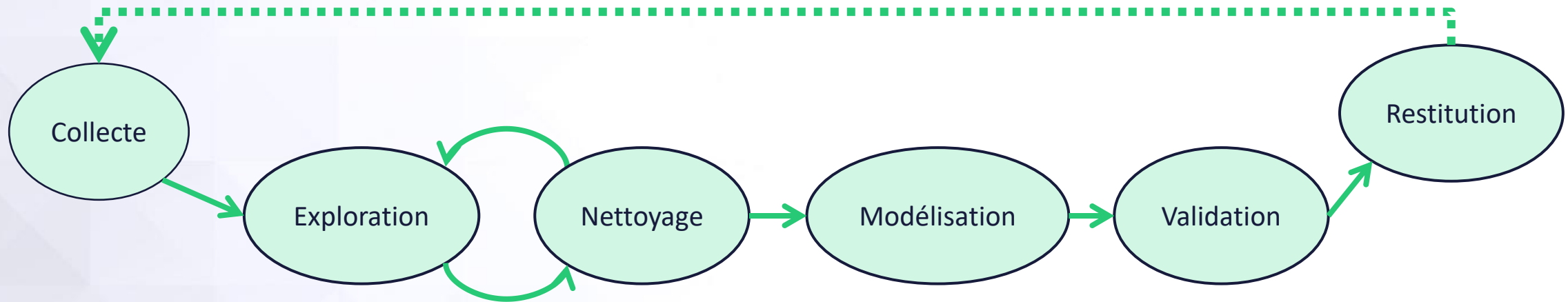
Nouveaux Outils



PROJET DE DATA SCIENCE

LES DIFFÉRENTES ÉTAPES

LES DIFFÉRENTES ÉTAPES D'UN PROJET DE DATA SCIENCE



Readr
Data.table

GGplot2
Plotly

Dplyr
Data.table

Caret

Rmarkdown
R Shiny



Pandas

Matplotlib
Seaborn
Plotly

Pandas

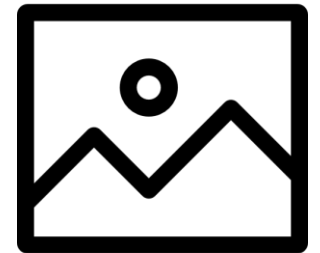
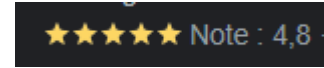
Scikit-learn

Jupyter
Streamlit / Dash
(Bokeh)

COLLECTE – LES DIFFÉRENTS TYPES DE DONNÉES

Il existe de différents types de données :

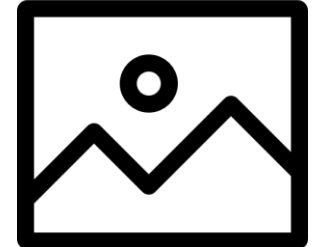
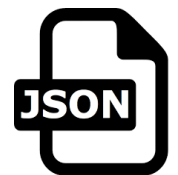
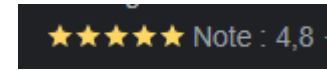
- Structurées
 - ...
- Semi-structurées
 - ...
- Non Structurées
 - ...



COLLECTE – LES DIFFÉRENTS TYPES DE DONNÉES

Il existe de différents types de données :

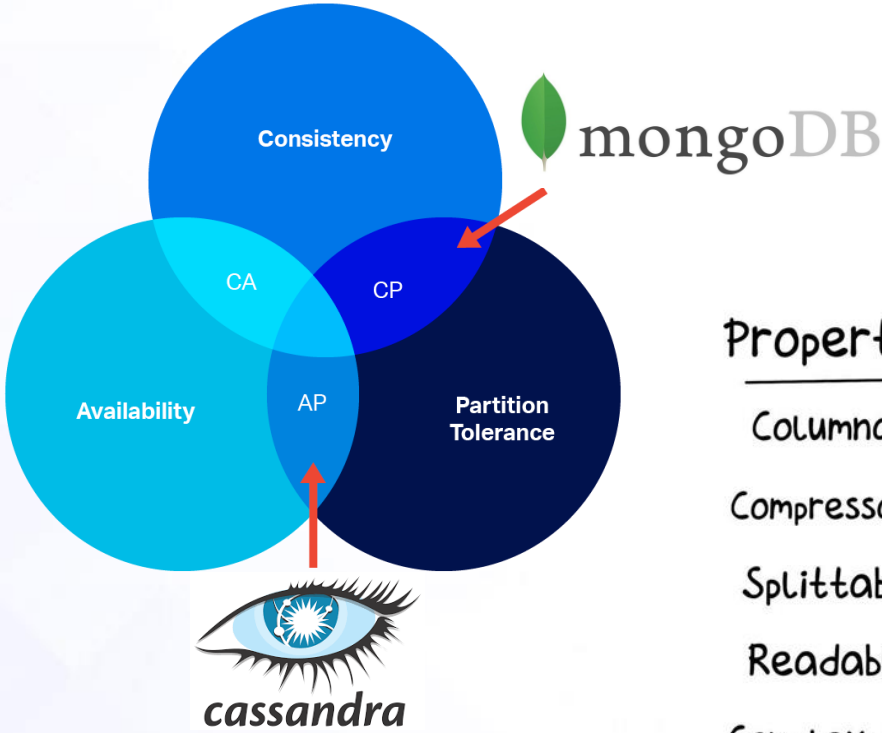
- Structurées
 - Base de données
 - Tableurs
 - Json
 - Xml
 - ...
- Semi-structurées
 - Pages web
 - Documents
 - ...
- Non Structurées
 - Texte
 - Images
 - Audio
 - Vidéo
 - ...





COLLECTE

Formats de données

- Fichiers
 - CSV
 - JSON
 - XML
 - Parquet vs Avro
- Base de données
 - SQL
 - NO-SQL
- API
- Web Scrapping
- Data WareHouse / Data Mart / DataLake / Data virtualisation
- Cloud



Properties	CSV	JSON	 Parquet	 AVRO
Columnar	✗	✗	✓	✗
Compressable	✓	✓	✓	✓
Splittable	✓*	✓*	✓	✓
Readable	✓	✓	✗	✗
Complex data structure	✗	✓	✓	✓
Schema evolution	✗	✗	✓	✓

@luminousmen.com



vs



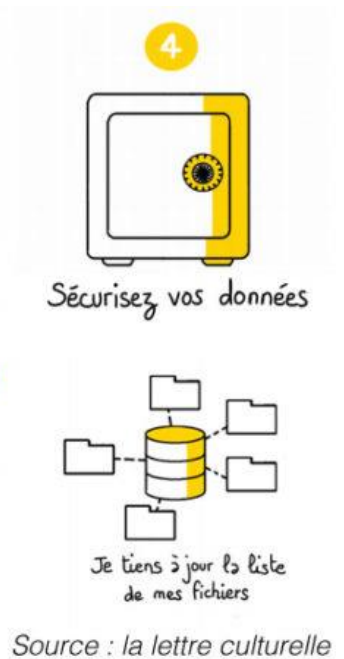
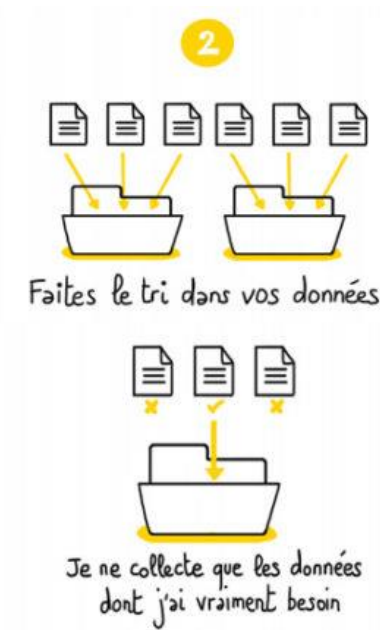
vs



COLLECTE

Étapes préliminaires à un projet :

- S'assurer que le projet est réalisable
 - Données disponibles
 - Données suffisantes
 - Données de qualité
- S'assurer que le projet est conforme au RGPD (voir avec le DPO)
- S'assurer d'avoir les ressources à disposition
 - Ressources
 - Délais
- (S'assurer que le projet est étiqe)
 - RSE
 - Biais
 - Utilité



CNIL

PROTÉGER les données personnelles
ACCOMPAGNER l'innovation
PRÉSERVER les libertés individuelles



**DATA
FOR
GOOD**



LESSWRONG

PROJET DE DATA SCIENCE

PRÉPARATION DES DONNÉES

EXPLORATION

Objectifs :

1. Comprendre les données
2. Vérifier la pertinence et la cohérence des données
3. Anticiper les problèmes
 - Valeurs manquantes
 - Doublons
 - Valeurs aberrantes
 - Biais
4. Création d'indicateurs
5. Visualiser



NETTOYAGE DES DONNÉES

OBJECTIF : AVOIR DES DONNÉES PROPRES, COHÉRENTES ET COMPLÈTES (LE PLUS POSSIBLE).

	Nom	Sexe	Ville	Code Postal	latitude	longitude	Age	Taille	Salaire	Client	Num fidélité	Code Concurrent
0	Paul	M	Paris	75000	48.856614	2.352222	18	1.80	1500.0	True	1235	None
1	Pierre	M	Nante	44000	47.218371	-1.553621	25	1.75	20000.0	False	None	A
2	Jacques	M	Lyon	69000	45.764043	4.835659	32	170.00	2500.0	True	1237	None
3	Julie	F	Paris	75000	48.856614	2.352222	45	1.65	NaN	False	None	None
4	Anne	F	Nantes	44000	47.218371	-1.553621	18	1.80	1500.0	True	1238	B
5	Marie	F	Lyon	69000	45.764043	4.835659	25	1.75	2000.0	False	None	None
6	Andrée	F	Paris	75000	48.856614	2.352222	322	1.70	2500.0	True	None	None
7	Fassou	M	Nantes	44000	47.218371	-1.553621	45	1.65	3000.0	False	None	None
8	James	M	Lyon	69000	45.764043	4.835659	18	1.80	1500.0	True	1240	None
9	Bob	M	Paris	75000	48.856614	2.352222	25	1.75	2000.0	False	None	None

NETTOYAGE DES DONNÉES

- Suppression d'individus
 - Outliers
 - Valeurs Manquantes
 - Doublons
 - ...
- Suppression de variables
 - Variables redondantes (ex : ville et code postal)
 - Variables non-pertinentes (ex : id, date de création, ...)
- Imputation / Modification de données
 - Imputation par une moyenne
 - Imputation multivariée (ex : KNN, MissForest ,)
 - Création d'une modalité valeur manquante
- Transformation de variables
 - Regroupement de modalités, ou de variables.
 - Standardisation ou Normalisation

	Nom	Sexe	Ville	Code Postal	latitude	longitude	Age	Taille	Salaire	Client	Num fidélité	Code Concurrent
0	Paul	M	Paris	75000	48.856614	2.352222	18	1.80	1500.0	True	1235	None
1	Pierre	M	Nante	44000	47.218371	-1.553621	25	1.75	20000.0	False	None	A
2	Jacques	M	Lyon	69000	45.764043	4.835659	32	170.00	2500.0	True	1237	None
3	Julie	F	Paris	75000	48.856614	2.352222	45	1.65	NaN	False	None	None
4	Anne	F	Nantes	44000	47.218371	-1.553621	18	1.80	1500.0	True	1238	B
5	Marie	F	Lyon	69000	45.764043	4.835659	25	1.75	2000.0	False	None	None
6	Andrée	F	Paris	75000	48.856614	2.352222	322	1.70	2500.0	True	None	None
7	Fassou	M	Nantes	44000	47.218371	-1.553621	45	1.65	3000.0	False	None	None
8	James	M	Lyon	69000	45.764043	4.835659	18	1.80	1500.0	True	1240	None
9	Bob	M	Paris	75000	48.856614	2.352222	25	1.75	2000.0	False	None	None

PROJET DE DATA SCIENCE **MODÉLISATION**

MODÉLISATION : 2 TYPES

- Apprentissage supervisé
 - Classification
Prédire une variable qualitative
Ex : arbre de décision, forêts aléatoires
 - Régression
Prédire une variable qualitative
Ex : régression linéaire , ...
- Apprentissage non-supervisé (clustering)
Regrouper des individus de façon cohérente
Ex : KNN, DBSCAN, ...

MODÉLISATION : 2+ TYPES

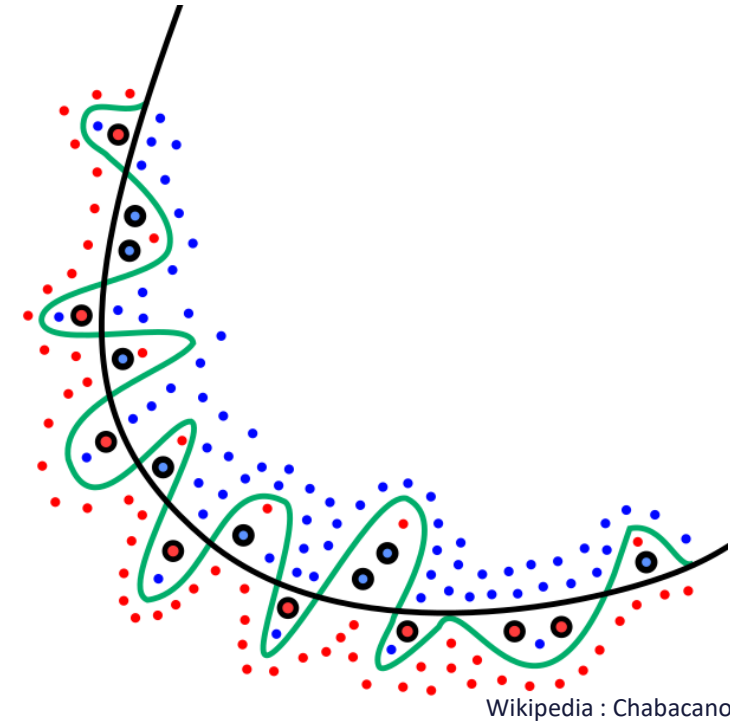
- Apprentissage supervisé
 - Classification
Prédire une variable qualitative
Ex : arbre de décision, forêts aléatoires
 - Régression
Prédire une variable qualitative
Ex : régression linéaire , ...
- Apprentissage non-supervisé (clustering)
Regrouper des individus de façon cohérente
Ex : KNN, DBSCAN, ...
- Apprentissage par renforcement
Apprendre à un agent à prendre des décisions
Ex : Q-Learning, ...
Applications : robotique, jeux, ..

MODÉLISATION : SUR-APPRENTISSAGE

Le modèle apprend par cœur les données d'entraînement et n'est pas capable de généraliser sur de nouvelles données

Causes possibles :

- Complexité du modèle
- Données d'entraînement trop peu représentatives
- Données d'entraînement trop peu nombreuses
- ...



MODÉLISATION : SUR-APPRENTISSAGE

Pour éviter de sélectionner un modèle qui souffre de sur-apprentissage, il faut évaluer les performances du modèle sur des données qu'il n'a jamais vu.

Validation Hold-Out :

On découpe les données en 2 (ou 3) parties : train, (validation) et test. On entraîne le modèle sur les données d'entraînement et on évalue les performances sur les données de test.

- train : données d'entraînement
- validation : données de validation, pour le réglage des hyperparamètres
- test : données de test, pour l'évaluation des performances

VALIDATION CROISÉE

Validation Croisée

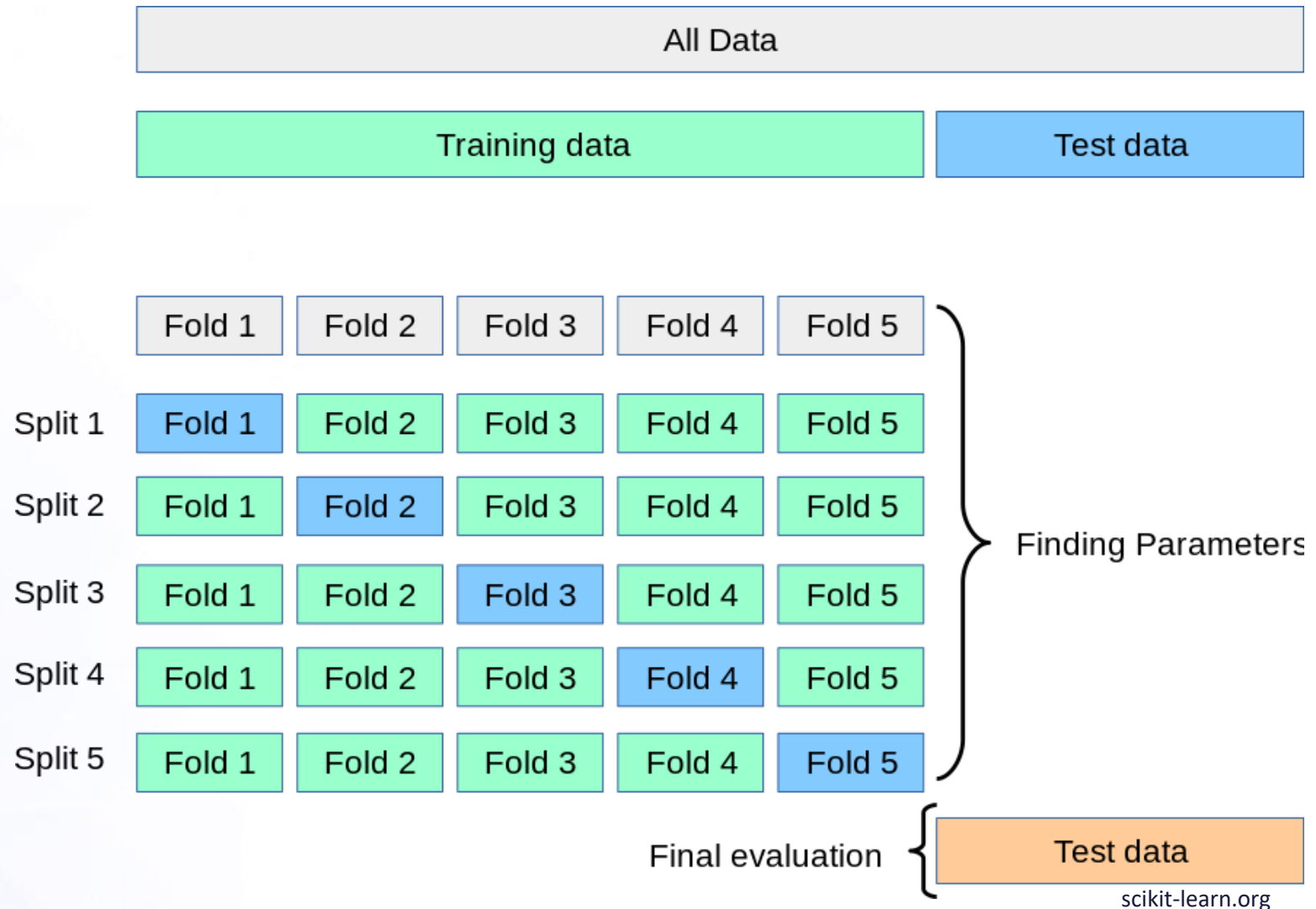
- Validation K-fold

On découpe les données en k parties égales.

On entraîne le modèle sur k-1 parties et on évalue les performances sur la partie restante.

On répète l'opération k fois en changeant à chaque fois la partie de test.

Le score de notre modèle sera la moyenne des k scores obtenus



VALIDATION CROISÉE

Validation Croisée

- Validation Leave one out

On entraîne le modèle sur $n-1$ données et on évalue les performances sur la donnée restante. On répète l'opération n fois en changeant à chaque fois la donnée de test.

- Validation Leave p out

On entraîne le modèle sur $n-p$ données et on évalue les performances sur les p données restantes. On répète l'opération n fois en changeant à chaque fois les données de test.

VALIDATION CROISÉE

La validation hold-out :

- Fonctionnement : séparation en deux blocs de la base de données (une partie d'apprentissage, une de test).
- Utilité : en présence de gros jeux de données, avec peu de temps disponible.
- Quelle proportion choisir ? 70% d'apprentissage, 30% de validation en général.

Y_1	X_1^1	...	X_1^p
Y_2	X_2^1	...	X_2^p
Y_3	X_3^1	...	X_3^p
Y_4	X_4^1	...	X_4^p
Y_5	X_5^1	...	X_5^p
Y_6	X_6^1	...	X_6^p

f_{D_A}

$$\begin{array}{l} \dots \\ \leftrightarrow l(f_{D_A}(X_4), Y_4) \\ \leftrightarrow l(f_{D_A}(X_5), Y_5) \\ \leftrightarrow l(f_{D_A}(X_6), Y_6) \end{array}$$

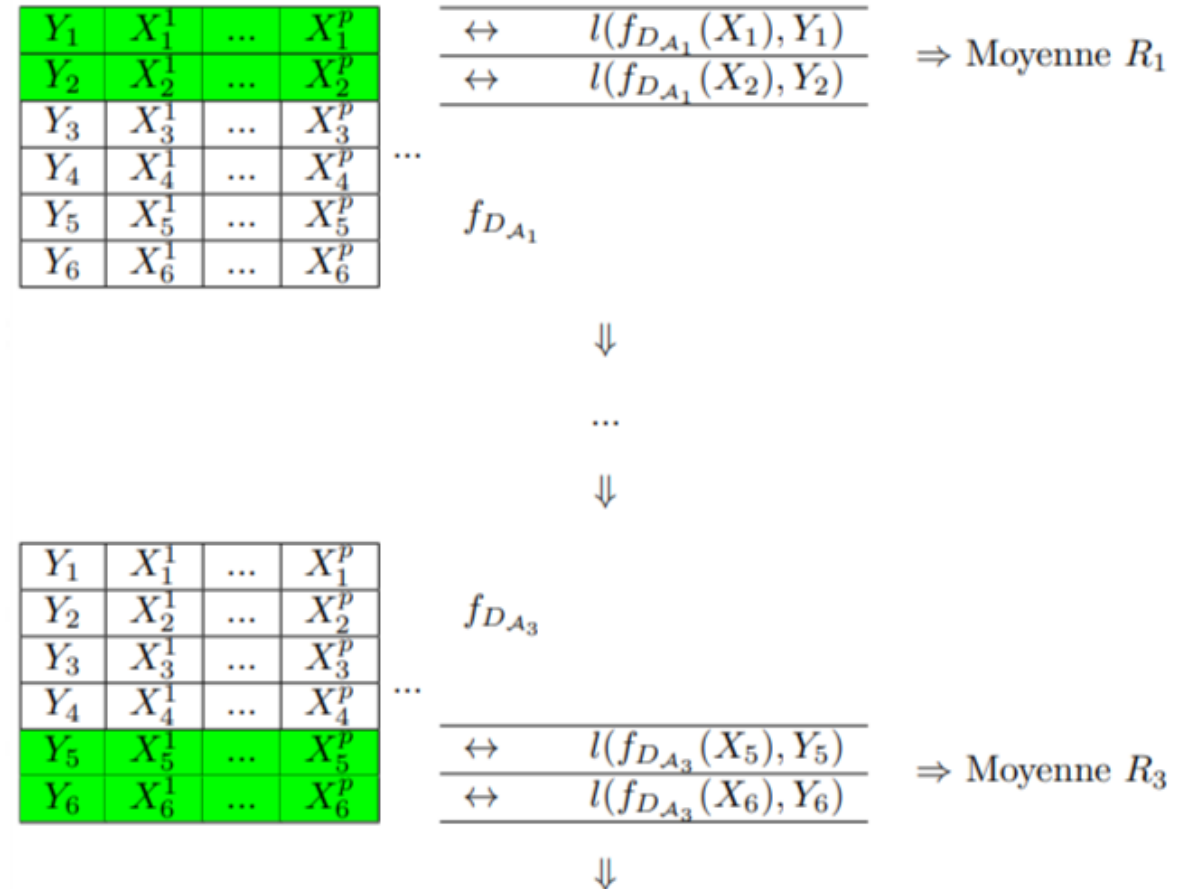
\Rightarrow

Moyenne des erreurs de prédiction :
 $\frac{1}{3} \sum_{i \in \mathcal{V}} l(Y_i, f_{D_A}(X_i))$

VALIDATION CROISÉE

La validation K-fold :

- Fonctionnement : séparation de la base en K échantillons. Apprentissage sur K-1 échantillons, test sur le dernier bloc.
- Utilité : en présence de gros jeux de données (meilleure approximation que la validation hold-out, mais plus longue).
- Quel K choisir ? K = 10 est recommandé.



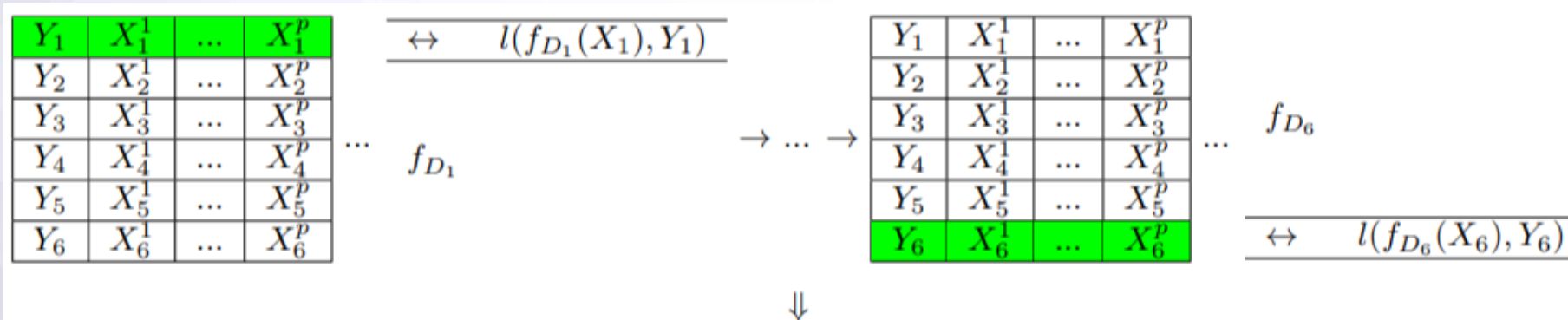
Moyenne des 3 moyennes :

$$\frac{1}{3} \sum_{i=1}^3 R_i$$

VALIDATION CROISÉE

La validation leave-one-out :

- Fonctionnement : cas spécifique de la méthode K-fold avec $p = n$.
- Utilité : en présence de petits jeux de données.



Moyenne des 6 erreurs de prédiction :

$$\frac{1}{6} \sum_{i=1}^6 l(Y_i, f_{D_A}(X_i))$$

VALIDATION – ECHANTILLONNAGE

RÉSUMÉ

Echantillonnage (Data sampling) | *package utile* : <https://imbalanced-learn.org/>

Objectif : Avoir des données plus équilibrées

Sur-échantillonnage (augmenter le nombre d'individu d'une classe minoritaire)

- Sur-échantillonnage aléatoire pour la classe minoritaire
 - Reproduction des données minoritaire aléatoirement
- Synthetic Minority Oversampling Technique (SMOTE)
 - Generation de données synthétiques
- ADASYN: Adaptive Synthetic Sampling
 - Génération de données synthétique selon en fonction de la densité des données

Sous-échantillonnage (Diminuer le nombre d'individus de la classe majoritaire)

- Sous-échantillonnage aléatoire pour la classe majoritaire
- ClusterCentroids
 - Regroupement d'individus avec des algorithmes de clustering et récupération des centroïdes
- NearMiss
 - Consiste à supprimer des données de la classe majoritaire en fonction de la distance avec les données de la classe minoritaire
- ...

VALIDATION – INDICATEUR DE PERFORMANCE

PROBLÈME DE RÉGRESSION

R^2

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

```
from sklearn.metrics import r2_score
y_pred = [3, -0.5, 2, 7]
y_true = [2.5, 0.0, 2, 8]
r2_score(y_true, y_pred)
```

Avantage : Permet de comparer plusieurs modèles entre eux même s'ils n'ont pas le même nombre de variables

Inconvénient : Sensible aux valeurs aberrantes

Mean Square Error

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

```
from sklearn.metrics import mean_squared_error
y_pred = [3, -0.5, 2, 7]
y_true = [2.5, 0.0, 2, 8]
Mean_squared_error(y_true, y_pred)
```

Avantage : Facile à interpréter

Inconvénient : Attention à l'échelle de la variable à prédire

VALIDATION – INDICATEURS DE PERFORMANCE

PROBLÈME DE CLASSIFICATION

Accuracy


$$\text{Accuracy} = \frac{\text{nb individus correctement prédit}}{\text{nb individus}}$$


Avantages :

- Facilement interprétable.

Inconvénients :

- A traiter avec précaution lors d'un déséquilibre de proportion de la variable à prédire.

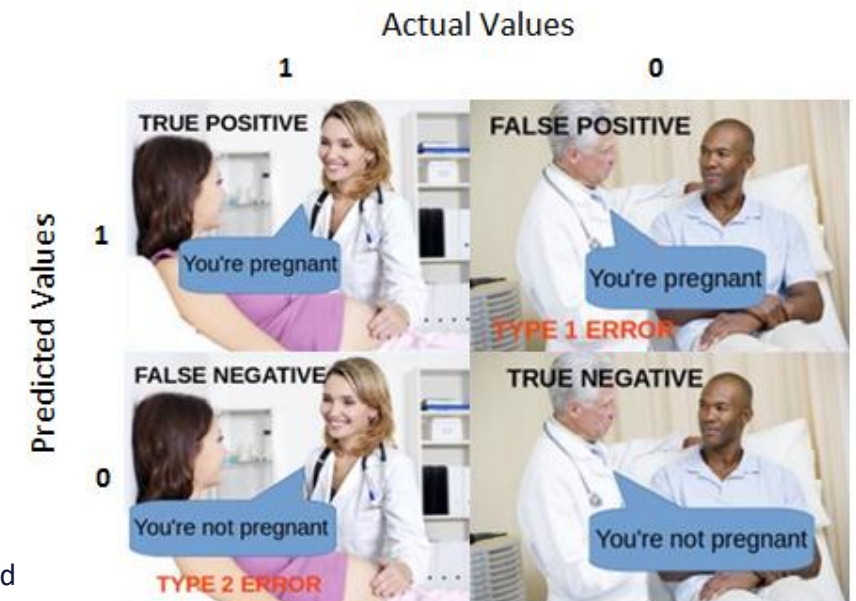
Yvrai : 

Ypred : 

```
from sklearn.metrics import accuracy_score  
print('Accuracy : ', accuracy_score(Yvrai, Ypred))
```

✓ 0.0s

Accuracy : 0.6666666666666666





VALIDATION – INDICATEURS DE PERFORMANCE

PROBLÈME DE CLASSIFICATION

Précision

$$precision = \frac{VP}{VP + FP}$$

Yvrai : 
Ypred : 

Matrice de confusion

Valeurs réelles	blue	red	
	6	1	
blue	6	1	
red	3	2	
Valeurs prédites			
		blue	red

```
from sklearn.metrics import precision_score
print('Precision : ',
      precision_score(Yvrai, Ypred, pos_label='red'),
      'pour la classe red')
print('Precision : ',
      precision_score(Yvrai, Ypred, pos_label='blue'),
      'pour la classe blue')
```

✓ 0.0s

Precision : 0.6666666666666666 pour la classe red
Precision : 0.6666666666666666 pour la classe blue

Permet d'identifier la proportion d'identifications positives correctement prédites.

Recall / Rappel

$$Recall = \frac{VP}{VP + FN}$$

```
from sklearn.metrics import recall_score
print('Rappel : ',
      recall_score(Yvrai, Ypred, pos_label='red'),
      'pour la classe red')
print('Rappel : ',
      recall_score(Yvrai, Ypred, pos_label='blue'),
      'pour la classe blue')
```

✓ 0.0s

Rappel : 0.4 pour la classe red
Rappel : 0.8571428571428571 pour la classe blue

Permet d'identifier la proportion de résultats positifs réellement identifiés.

VALIDATION – INDICATEURS DE PERFORMANCE



PROBLÈME DE CLASSIFICATION

F_β score

$$F_\beta \text{ score} = (1 + \beta^2) * \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}$$

Où beta représente combien on accorde d'importance au recall par rapport à la précision

(1 par défaut = on accorde autant d'importance aux deux indicateurs)

Yvrai : 
Ypred : 

```
from sklearn.metrics import f1_score

print('F1 : ',
      f1_score(Yvrai, Ypred, pos_label='red'),
      'pour la classe red')
print('F1 : ',
      f1_score(Yvrai, Ypred, pos_label='blue'),
      'pour la classe blue')
```



✓ 0.0s

F1 : 0.5 pour la classe red

F1 : 0.75 pour la classe blue

VALIDATION – INDICATEUR DE PERFORMANCE

RÉSUMÉ

Yvrai : 
Ypred : 

Matrice de confusion

Valeurs réelles	blue	6 VP	1 FP
	red	3 FN	2 VN
		blue	red
		Valeurs prédites	

$Recall = \frac{VP}{VP+FN}$ On s'intéresse aux prédictions positives.

Recall de 1 : on ne prédit aucun faux positif

Ex : sur 100 prédictions positives, combien sont réellement positives ?

$Precision = \frac{VP}{VP+FP}$ On s'intéresse aux observations positives.

Précision de 1 : on ne prédit correctement tous les vrais positifs

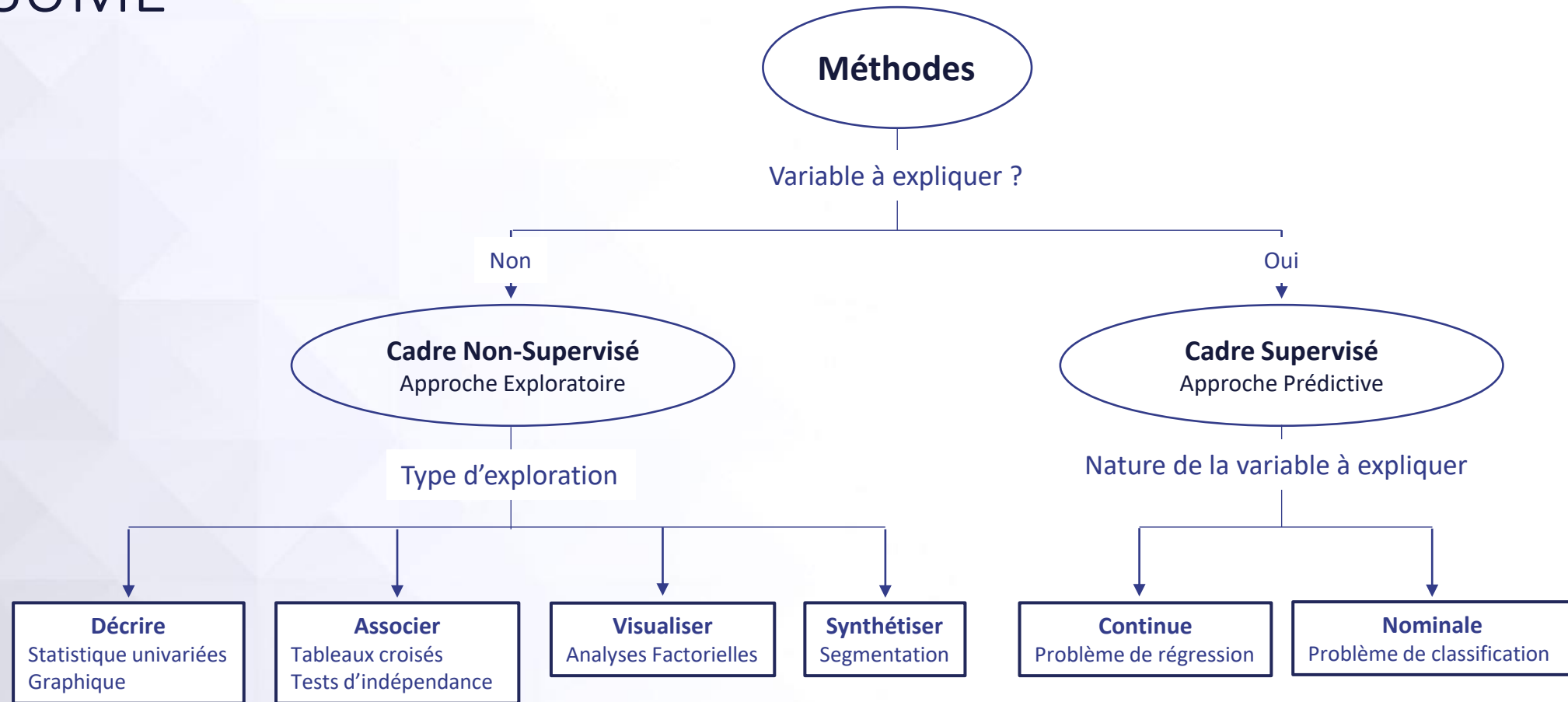
Ex : sur 100 individus positifs, combien sont détectés ?

$Accuracy = \frac{VP+VN}{VP+FP+FN+VN}$ On s'intéresse ici à la globalité des prédictions.

Accuracy de 1 : Le modèle prédit parfaitement les données.

Ex : sur 100 prédictions, combien sont correctes ?

RÉSUMÉ



TP1 – EXPLORATION ET NETTOYAGE DES DONNÉES