

Projet de fin de module: *Data mining*

Sujet : Prévision du désabonnement de clients dans le secteur de télécommunication

Présenté par :

□ OUNACHAD Walid

Responsable :

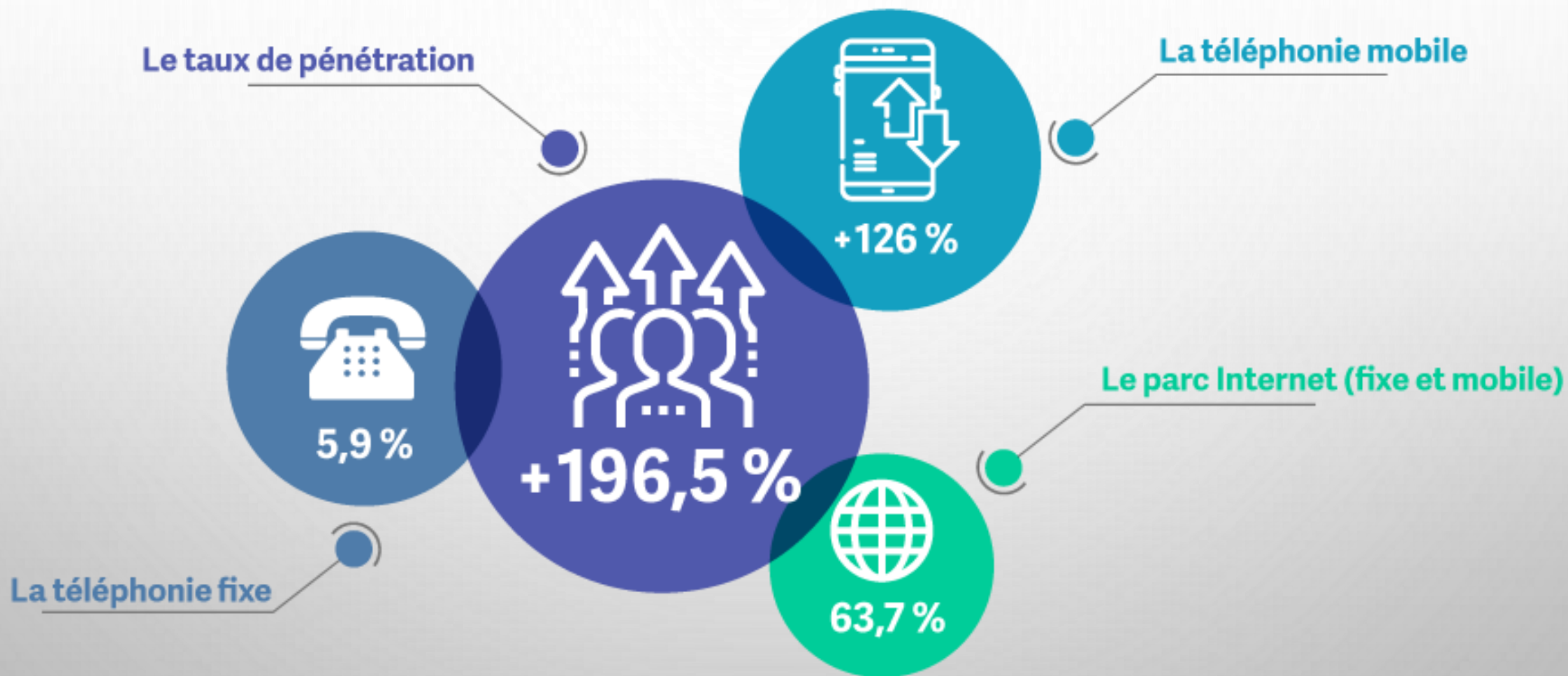
□ Pr:Sabiri

PLAN

- Introduction
- Etude de l'existant
- Solution proposée
- Mise en place
- Conclusion et perspectives

INTRODUCTION

INTRODUCTION



Le secteur des télécommunications poursuit sa transformation.

C'est ce que montrent les chiffres publiés par l'ANRT décrivant la situation du secteur au Maroc, à fin décembre 2017.⁴

Introduction



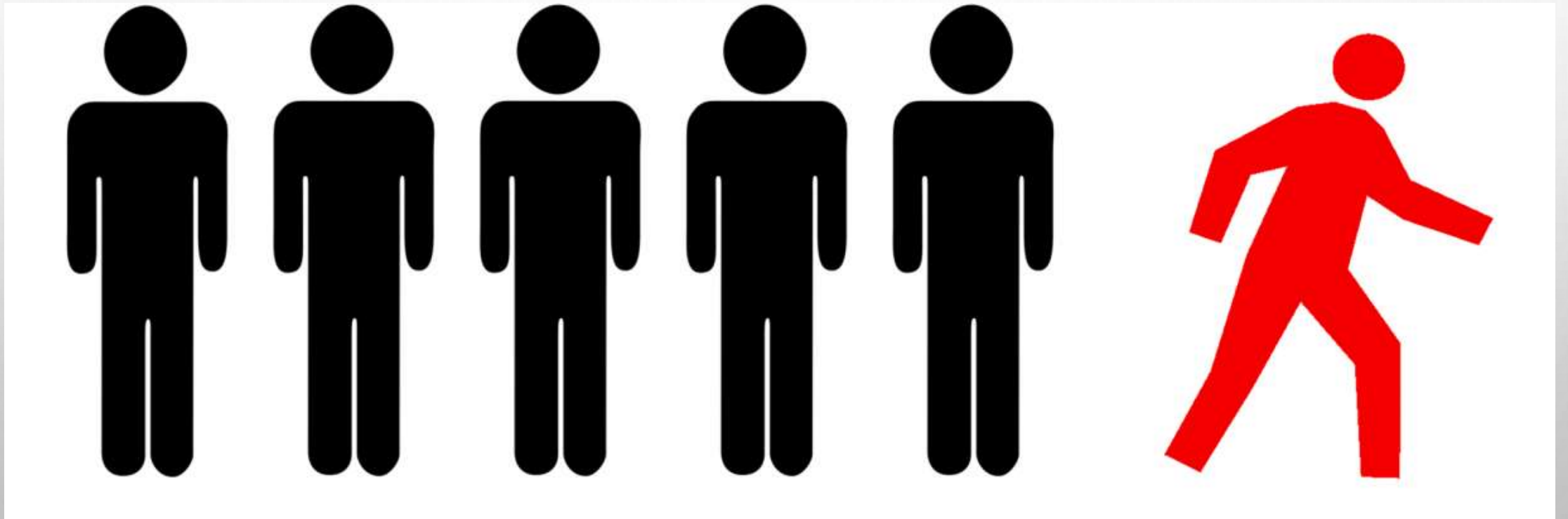
Saturation

INTRODUCTION



Saturation
=
Clients libres

Introduction



CHURN

INTRODUCTION

Définition :

Churn : (Du français [Attrition])

- Exprime le taux de de perdition de clients pour une entreprise

Les coûts d'acquisition d'un nouveau client.

5X

Les coûts de maintien d'un client actuel



INTRODUCTION

- Donc maintenant la question qui se pose:
- Pourquoi les clients se désabonnent, et comment pouvons-nous les prédire?



ETUDE DE L'EXISTANT

INTRODUCTION

- L'objectif des études sur le churn est la détection des individus qui ont l'intention de quitter l'organisation afin d'améliorer la prise de décision et de mettre en place des actions de rétention.
- La problématique est donc la même pour ces deux phénomènes qui sont souvent analysés à l'aide de techniques prédictives similaires du data mining.

LE DATA MINING, C'EST QUOI ?

- Le Data Mining (littéralement « forage des données »)
- Couvre l'ensemble des outils et méthodes qui permettent d'extraire des connaissances à partir de grandes quantités de données.
- Grâce à des méthodes d'analyse de données et de statistiques exploratoires, on pratique le Data Mining depuis plus de 30 ans dans de nombreux secteurs d'activités. Mais si le concept est particulièrement à la mode aujourd'hui, c'est que **certaines évolutions récentes méritent d'être soulignées.**

LE DATA MINING, C'EST QUOI ?

1. **Big Data : la volumétrie des données et leur variété suivent une croissance exponentielle.**

L'omniprésence d'Internet, la numérisation croissante des interactions de tout type, l'Open Data et l'émergence des objets connectés génèrent des volumes de données de plus en plus vertigineux. La variété des données s'est accrue aussi, avec notamment le développement des réseaux sociaux : images, sons, textes sont désormais également des données à analyser.

2. **Machine Learning : de nouvelles techniques d'analyse permettent de traiter rapidement ces grands volumes de données.**

Venues de l'univers de l'intelligence artificielle, ces nouvelles techniques enrichissent la boîte à outils des data analysts.

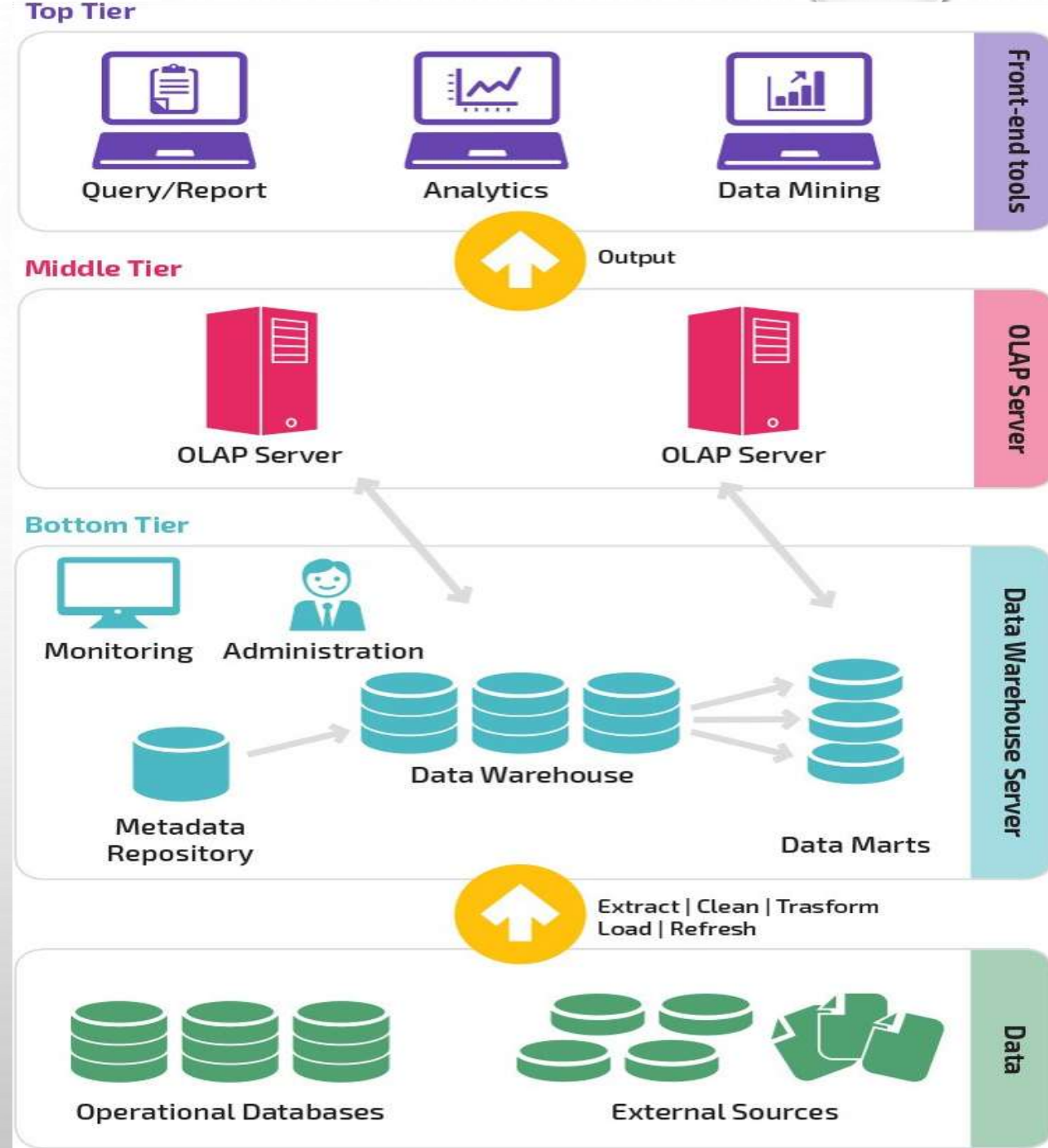
3. **Fiabilité et objectivité des données internes.**

Les méthodes statistiques traditionnelles analysaient surtout des données provenant d'enquêtes ou de sondages. Le Data Mining se concentre lui sur les **données internes** qui circulent dans les tuyaux informatiques des entreprises et notamment les données clients : informations issues des cartes de fidélité, transactions, etc.

LE DATA MINING : POUR QUOI FAIRE ?

- Le Data Mining permet à une entreprise de :
- **1. Mieux comprendre ses clients** en établissant un profil très fin de chacun d'entre eux, anticiper l'évolution de leurs besoins et ainsi adapter ainsi sa politique de fidélisation ;
- **2. Découvrir des niches rentables** nécessitant un traitement marketing spécifique ;
- **3. Optimiser l'adéquation de son offre à chaque profil** : adapter sa politique commerciale et sa tarification aux différents profils de clientèle, adapter ses canaux de distribution et sa communication aux différents profils, optimiser l'impact et la rentabilité des offres commerciales ;
- **4. Cibler finement ses actions** de recrutement sur des prospects à fort potentiel et plus généralement cibler finement ses actions de développement : cross-selling, up-selling ;
- **5. Adapter le calendrier de ses actions marketing aux différents profils.**
- **Pour conclure et pour résumer, le Data Mining est le process qui conduit de la Big Data à la Smart Data.**

Architecture typique d'un système de Data mining



LE DATA MINING : QUELLES DONNÉES SÉLECTIONNE-T-ON ?

EXEMPLE DE DONNEES CROISEES

#DataMining

VARIABLES NOMINATIVES

Nom, prénom, raison sociale, SIRET, adresse(s) physique(s), numéro(s) de téléphone, e-mail(s)

DONNEES SOCIO-DEMOGRAPHIQUES

Données personnelles, familiales, professionnelles, patrimoniales, géographiques

COMPORTEMENT D'ACHATS

Historiques des achats, modes d'achats, modes de paiement, montants, produits et contrats, devis...

DONNEES RELATIONNELLES

Historique des messages sortants et entrants, actions/réactions aux campagnes de sollicitation, demande d'informations, SAV, canaux privilégiés (recrutement/contact, commande, livraison...)

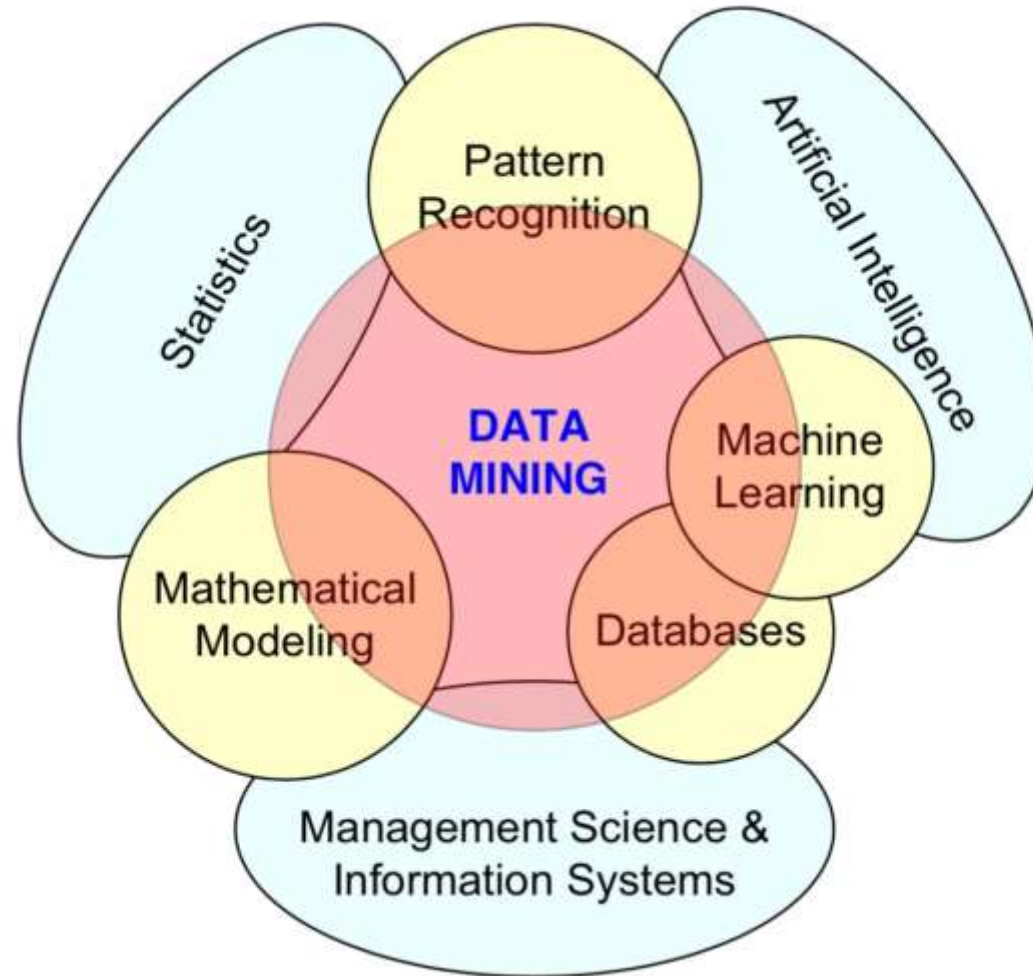
COMPORTEMENT DE NAVIGATION

Origine des visites, mots clés, pages visitées, taux de rebond, taux de conversion, parcours, fréquences des visites, activités sur les réseaux sociaux...

DONNEES ETUDES INTERNES

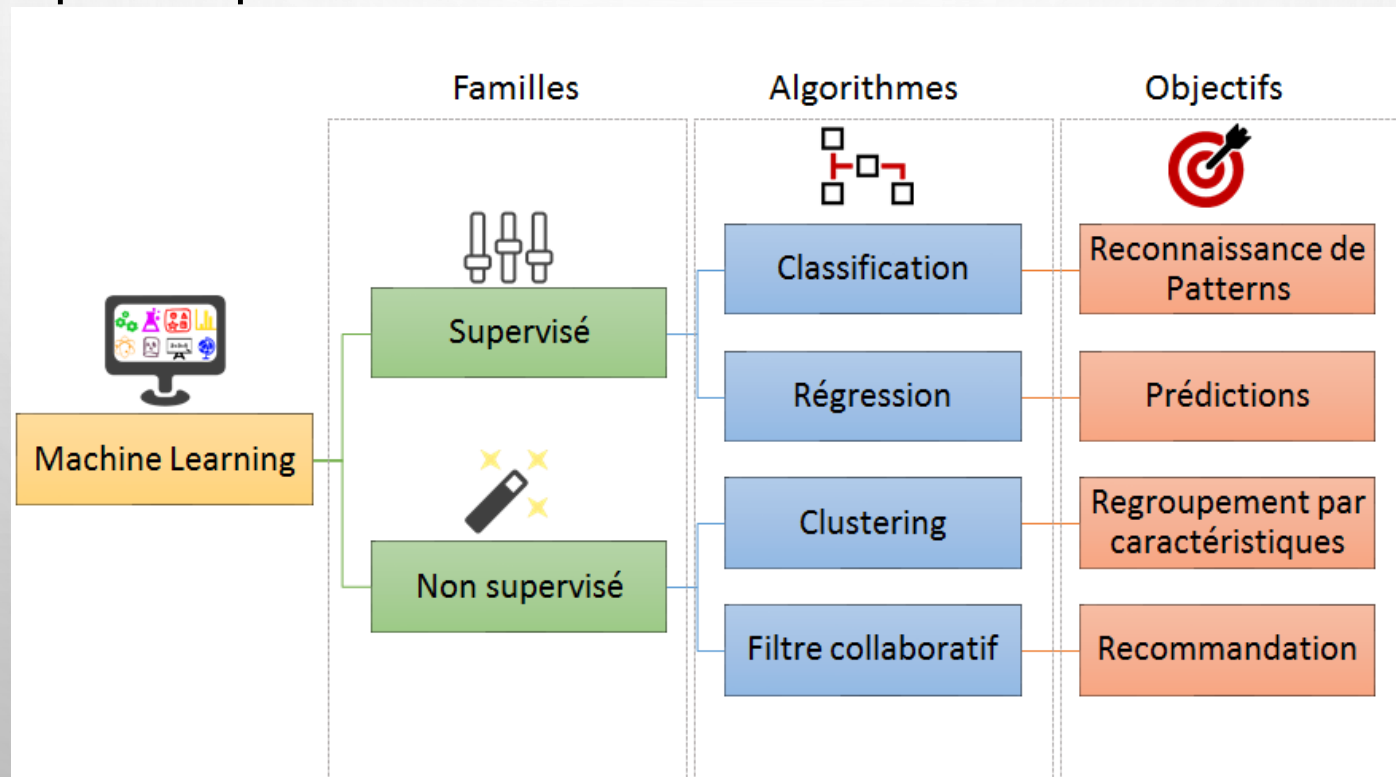
Données issues des études internes : enquêtes ad-hoc, profil client, segmentation(s), scores(s)

DATA MINING : ASPECT PLURIDISCIPLINAIRE

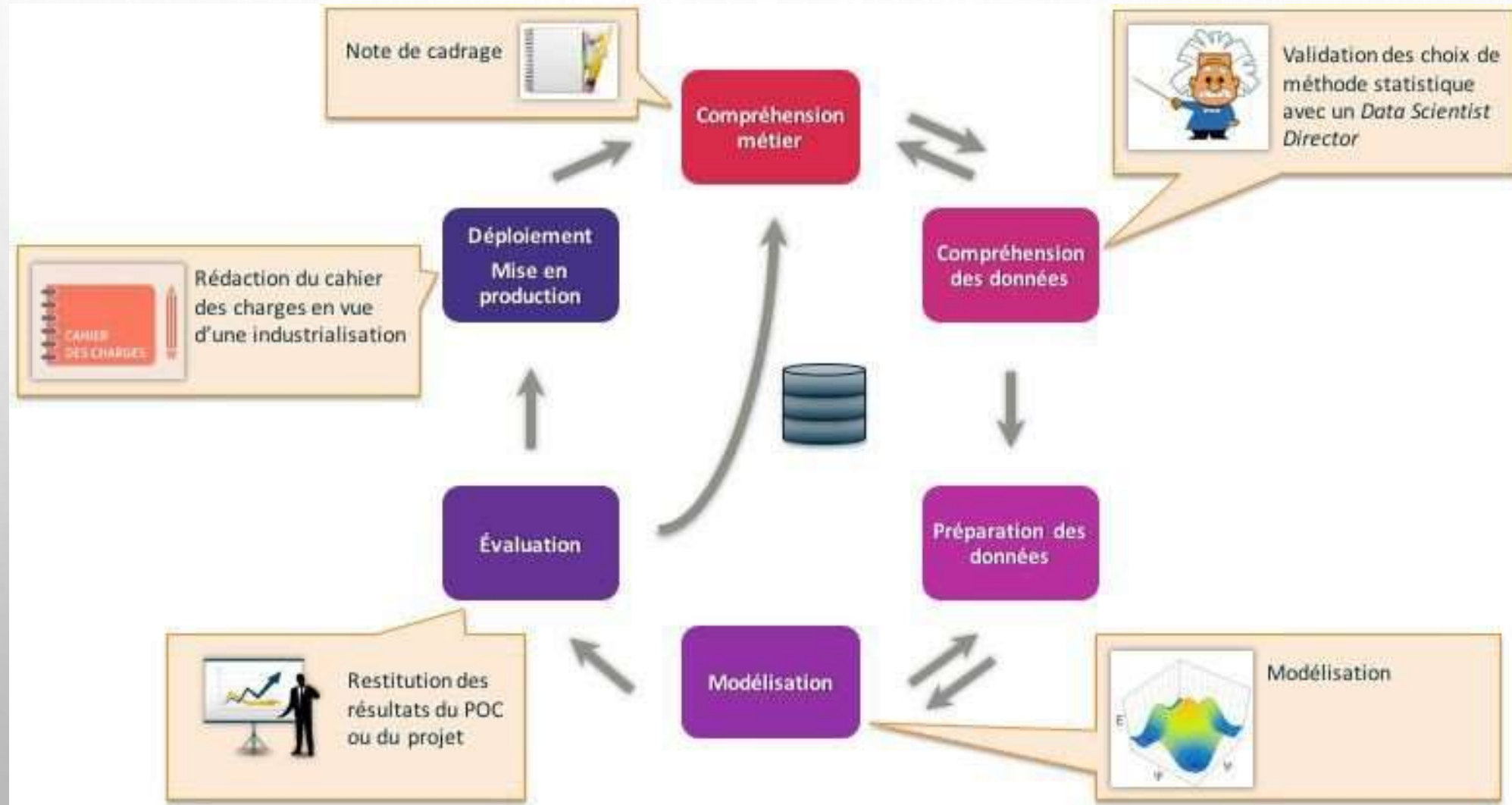


MACHINE LEARNING

- Le **Machine Learning** et l'IA (Intelligence Artificielle) font désormais partie intégrante du paysage technologique.
- Le premier utilise toute une série d'algorithmes éprouvés dont voici les principaux représentants :



DATA MINING: CRISP-DM CYCLE DE VIE



SOLUTION PROPOSÉE

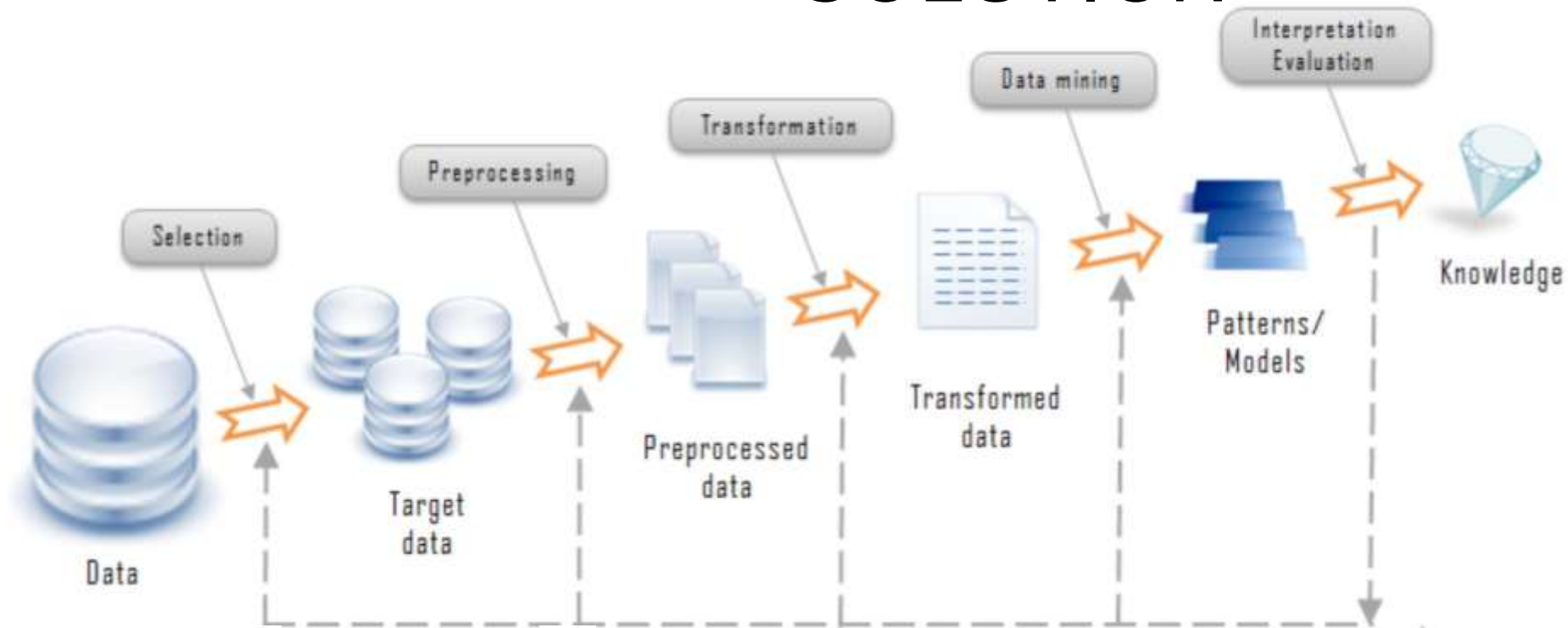
SOLUTION PROPOSÉE

- L'objectif de ce projet est de développer une méthode de prédiction du churn dans le secteur des télécoms. Le résultat attendu est un algorithme qui identifie les clients les plus proches d'abandonner leur opérateur actuel.
- Bien que de nombreuses approches aient été menées au cours des dernières années, il existe encore de nombreuses possibilités d'améliorer le travail actuel dans ce domaine. Les données qui seront utilisées lors de ce travail sont fondamentales pour assurer la qualité de la solution finale.

CHOIX TECHNOLOGIQUES

- **Distribution Big Data** : Cloudera
- **Datawarehouse** : Hive
- **Technologie de programmation** : Python, SQL
- **Bibliothèques utilisées** : Spark SQL, Numpy, Pandas, Sci-kit learn
- **DataViz** : Matplotlib, Searborn

CONCEPTION ET ARCHITECTURE DE LA SOLUTION

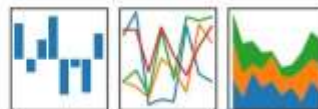


Spark SQL



pandas

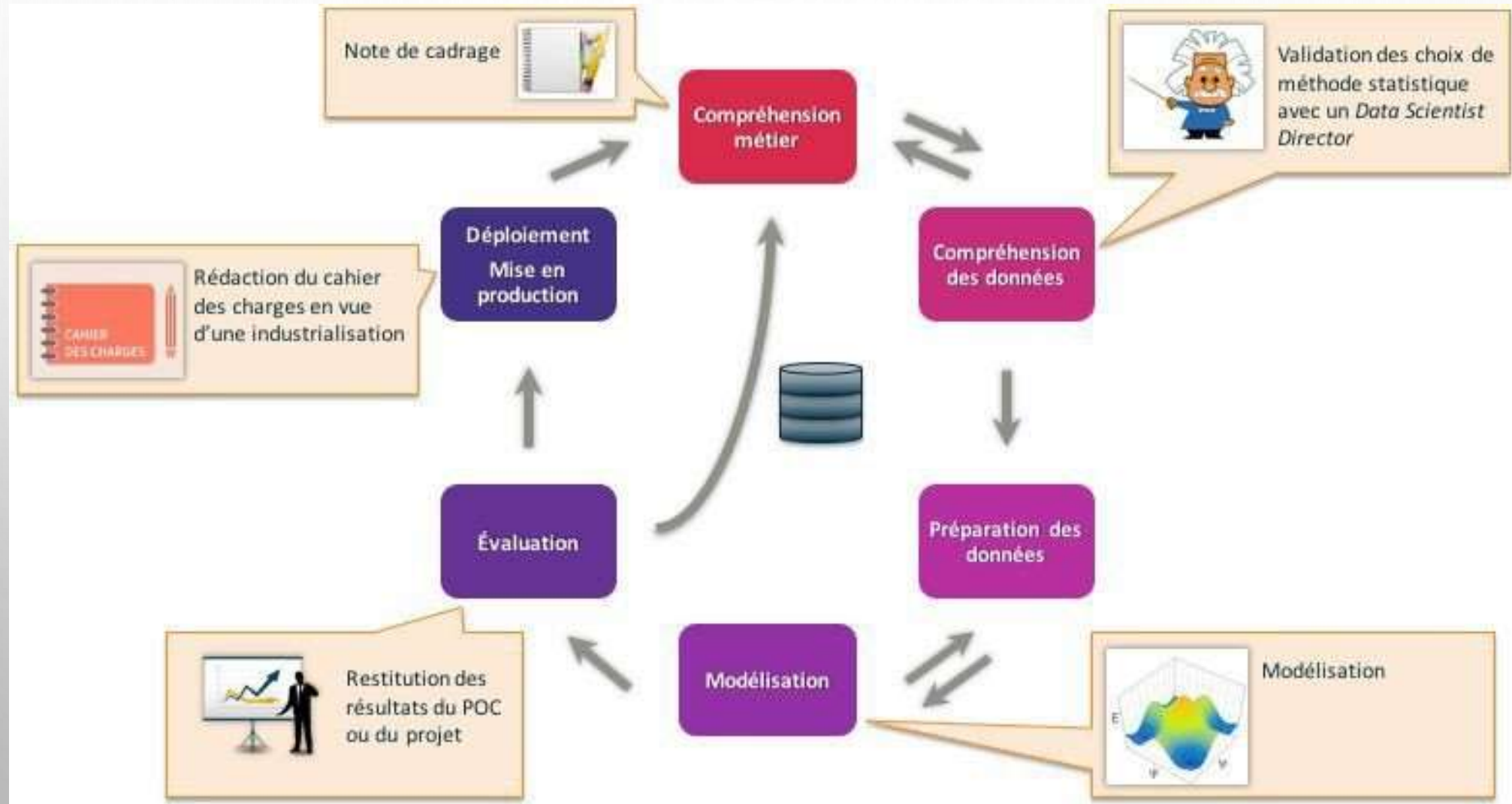
$$y_{it} = \beta^T x_{it} + \mu_i + \epsilon_{it}$$



Seaborn

matplotlib

CYCLE DE VIE DU PROJET



MISE EN PLACE

COMPRÉHENSION DU PROBLÈME MÉTIER

- **La perte de la clientèle ou d'abonnés** est toujours un problème grave pour l'industrie des télécommunications, car les clients n'hésitent pas à désabonner ou de changer l'opérateur, s'ils ne trouvent pas ce qu'ils recherchent. Les clients veulent certainement des prix compétitifs, et de la valeur ajoutée pour l'argent qu'ils payent et surtout, un service de haute qualité.
- **Le churn ou bien la perte de clientèle** est directement lié à **la satisfaction du client**. C'est un fait connu que le coût de l'acquisition d'un client est **beaucoup plus élevé** que le coût de la fidélisation d'un client, ce qui fait de la rétention des clients un prototype d'entreprise crucial. Il n'existe pas de modèle standard permettant de résoudre avec précision les problèmes de clientèle des fournisseurs de services télécoms mondiaux.
- **L'analyse Big Data avec le Machine Learning et Data mining** s'est révélée être un moyen efficace d'identifier et prédire le désabonnement des clients. Afin d'anticiper la rupture des clients avant qu'elle se produise.

COMPRÉHENSION DES DONNÉES

Name	Description	Value Type	Statistical Type
State	State abbreviation (like KS = Kansas)	String	Categorical
Account length	How long the client has been with the company	Numerical	Quantitative
Area code	Phone number prefix	Numerical	Categorical
International plan	International plan (on/off)	String, "Yes"/"No"	Categorical/Binary
Voice mail plan	Voicemail (on/off)	String, "Yes"/"No"	Categorical/Binary
Number vmail messages	Number of voicemail messages	Numerical	Quantitative
Total day minutes	Total duration of daytime calls	Numerical	Quantitative
Total day calls	Total number of daytime calls	Numerical	Quantitative
Total day charge	Total charge for daytime services	Numerical	Quantitative
Total eve minutes	Total duration of evening calls	Numerical	Quantitative
Total eve calls	Total number of evening calls	Numerical	Quantitative
Total eve charge	Total charge for evening services	Numerical	Quantitative
Total night minutes	Total duration of nighttime calls	Numerical	Quantitative
Total night calls	Total number of nighttime calls	Numerical	Quantitative
Total night charge	Total charge for nighttime services	Numerical	Quantitative
Total intl minutes	Total duration of international calls	Numerical	Quantitative
Total intl calls	Total number of international calls	Numerical	Quantitative
Total intl charge	Total charge for international calls	Numerical	Quantitative



Ce dataset téléchargé depuis le site de Kaggle:
Il contient 5000 lignes.
Et 21 variables dont
le variable cible: **Churn**

COMPRÉHENSION DES DONNÉES

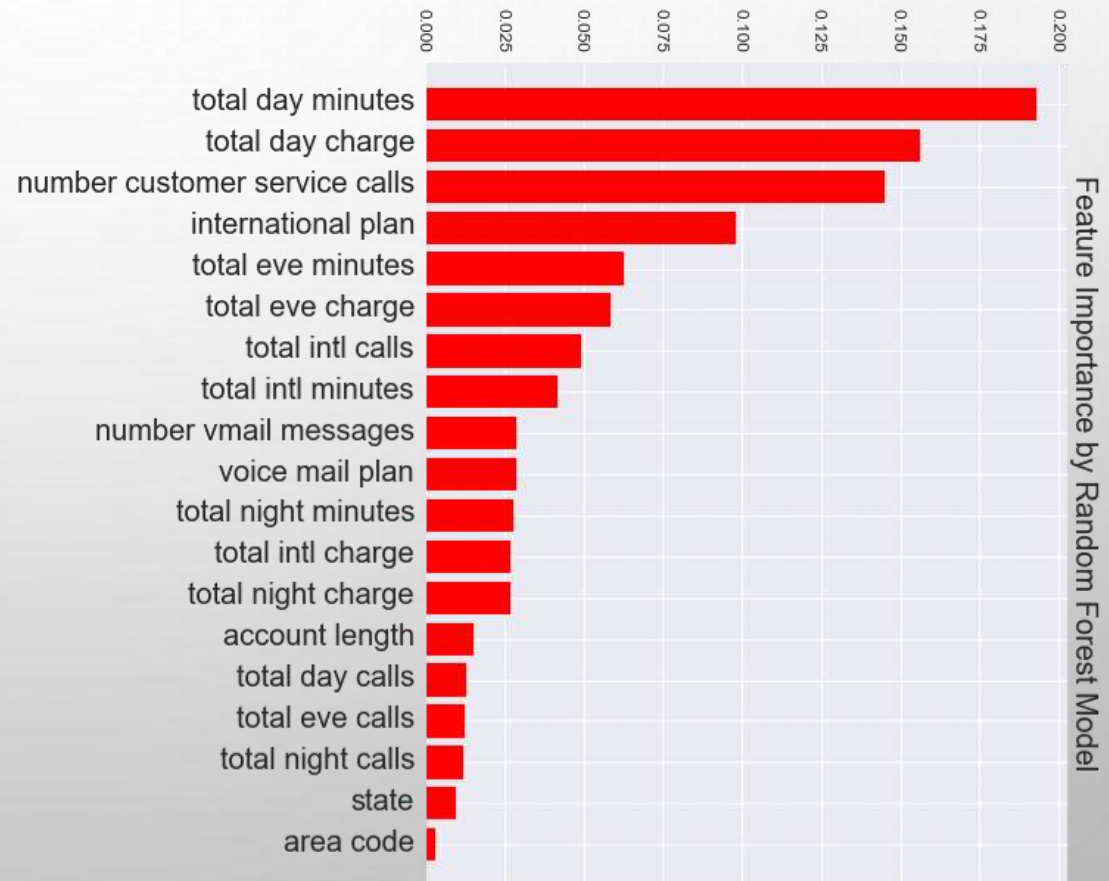
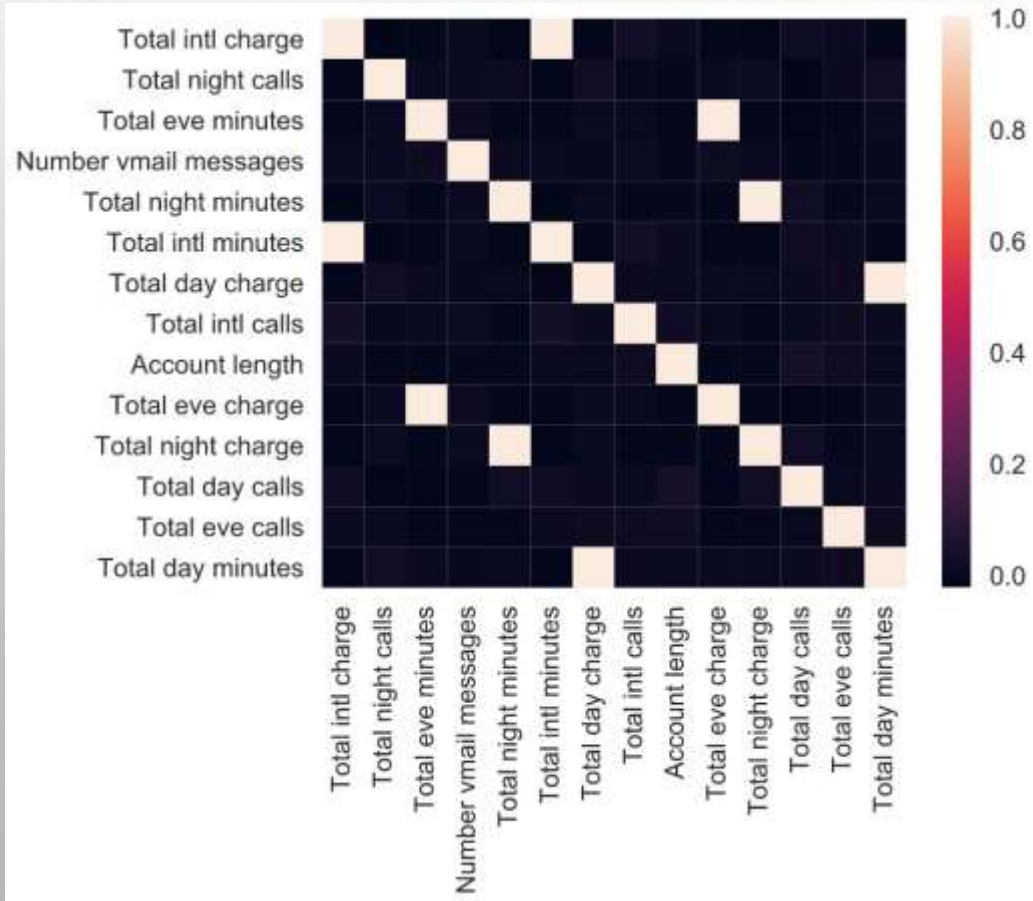
- Un aperçu sur l'ensemble des données

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	...	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Customer service calls	Churn
0	KS	128	415	No	Yes	25	...	11.01	10.0	3	2.70	1	False
1	OH	107	415	No	Yes	26	...	11.45	13.7	3	3.70	1	False
2	NJ	137	415	No	No	0	...	7.32	12.2	5	3.29	0	False
3	OH	84	408	Yes	No	0	...	8.86	6.6	7	1.78	2	False
4	OK	75	415	Yes	No	0	...	8.41	10.1	3	2.73	3	False

5 rows × 20 columns

COMPRÉHENSION DES DONNÉES

- Importance des variables les plus prédictives :



LA PRÉPARATION DES DONNÉES

- Vérification des valeurs nulles pour l'imputation des données :


Remarques:

- Aucune valeur nulle, donc pas besoin d'imputation.

```
id 0
state 0
account length 0
area code 0
phone number 0
international plan 0
voice mail plan 0
number vmail messages 0
total day minutes 0
total day calls 0
total day charge 0
total eve minutes 0
total eve calls 0
total eve charge 0
total night minutes 0
total night calls 0
total night charge 0
total intl minutes 0
total intl calls 0
total intl charge 0
number customer service calls 0
churn 0
dtype: int64
```

LA PRÉPARATION DES DONNÉES

- Certains algorithmes peuvent fonctionner directement avec des données catégorielles.
- Par exemple, **un arbre de décision (Decision tree)** peut être appris directement à partir de données catégorielles sans transformation de données requise (cela dépend de l'implémentation spécifique).
- De nombreux algorithmes d'apprentissage automatique ne peuvent pas fonctionner directement sur les données catégorielles. Ils exigent que toutes les variables d'entrée et les variables de sortie soient numériques.



	churn	internationalplan	voicemailplan
0	False	no	yes
1	False	no	yes
2	False	no	no
3	False	yes	no
4	False	yes	no

	churn	internationalplan	voicemailplan
0	0	0	1
1	0	0	1
2	0	0	0
3	0	1	0
4	0	1	0

LA MODÉLISATION

- **5 Algorithmes de Machine Learning les plus utilisés pour l'analyse du churn :**
 1. Machine à vecteurs de support (SVM)
 2. Méthode des k plus proches voisins (KNN)
 3. Arbre de decision (Decision Tree)
 4. Forêt d'arbres décisionnels (Random Forest)
 5. Régression logistique (Logistic Regression)

IMPLÉMENTATION DU MODÈLE : SVM

```
#Importation des librairies

from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split
data_train, data_test, label_train, label_test = train_test_split(x, y, test_size = 0.2, random_state = 42)

# affecter X en tant que DataFrame des entités et y en tant que série de la variable de résultat

X = df.drop('churn', axis = 1)
y = df.churn

# Support Vector Machine

from sklearn.svm import SVC
svm = SVC()
svm.fit(data_train, label_train)
svm_score_train = svm.score(data_train, label_train)
svm_score_test = svm.score(data_test, label_test)
svm_auc_score_train = roc_auc_score(label_train, svm.predict(data_train))
svm_auc_score_test = roc_auc_score(label_test, svm.predict(data_test)) |
```


IMPLÉMENTATION DU MODÈLE : KNN

```
#Importation des librairies
```

```
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split
data_train, data_test, label_train, label_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

```
# affecter X en tant que DataFrame des entités et y en tant que série de la variable de résultat
```

```
X = df.drop('churn', axis = 1)
y = df.churn
```

```
# K-Nearest Neighbours
```

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(data_train, label_train)
knn_score_train = knn.score(data_train, label_train)
knn_score_test = knn.score(data_test, label_test)
knn_auc_score_train = roc_auc_score(label_train, knn.predict(data_train))
knn_auc_score_test = roc_auc_score(label_test, knn.predict(data_test))
```

IMPLÉMENTATION DU MODÈLE : DECISION TREE

```
#Importation des librairies
```

```
from sklearn.metrics import roc_auc_score
```

```
from sklearn.model_selection import train_test_split
```

```
data_train, data_test, label_train, label_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

```
# affecter X en tant que DataFrame des entités et y en tant que série de la variable de résultat
```

```
X = df.drop('churn', axis = 1)
```

```
y = df.churn
```

```
# Decision Tree
```

```
from sklearn import tree
```

```
dt = tree.DecisionTreeClassifier()
```

```
dt.fit(data_train, label_train)
```

```
dt_score_train = dt.score(data_train, label_train)
```

```
dt_score_test = dt.score(data_test, label_test)
```

```
dt_auc_score_train = roc_auc_score(label_train, dt.predict(data_train))
```

```
dt_auc_score_test = roc_auc_score(label_test, dt.predict(data_test))
```


IMPLÉMENTATION DU MODÈLE : FORÊT D'ARBRES DÉCISIONNELS (RANDOM FOREST)

```
#Importation des librairies
```

```
from sklearn.metrics import roc_auc_score  
from sklearn.model_selection import train_test_split  
data_train, data_test, label_train, label_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

```
# affecter X en tant que DataFrame des entités et y en tant que série de la variable de résultat
```

```
X = df.drop('churn', axis = 1)  
y = df.churn
```

```
# Random Forest
```

```
from sklearn.ensemble import RandomForestClassifier  
rfc = RandomForestClassifier()  
rfc.fit(data_train, label_train)  
rfc_score_train = rfc.score(data_train, label_train)  
rfc_score_test = rfc.score(data_test, label_test)  
rfc_auc_score_train = roc_auc_score(label_train, rfc.predict(data_train))  
rfc_auc_score_test = roc_auc_score(label_test, rfc.predict(data_test))
```

IMPLÉMENTATION DU MODÈLE : RÉGRESSION LOGISTIQUE

```
#Importation des librairies
```

```
from sklearn.metrics import roc_auc_score  
from sklearn.model_selection import train_test_split  
data_train, data_test, label_train, label_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

```
# affecter X en tant que DataFrame des entités et y en tant que série de la variable de résultat
```

```
X = df.drop('churn', axis = 1)  
y = df.churn
```

```
# Logistic Regression
```

```
from sklearn.linear_model import LogisticRegression  
log = LogisticRegression()  
log.fit(data_train, label_train)  
log_score_train = log.score(data_train, label_train)  
log_score_test = log.score(data_test, label_test)  
log_auc_score_train = roc_auc_score(label_train, log.predict(data_train))  
log_auc_score_test = roc_auc_score(label_test, log.predict(data_test))
```

L'ÉVALUATION

- Les modèles mis en œuvre ont déjà été décrits et analysés séparément, mais l'objectif principal de cette recherche était d'identifier quel modèle peut être considéré comme le meilleur pour résoudre ce problème.
- Les modèles mis en œuvre ont déjà été décrits et analysés séparément, mais l'objectif principal de cette recherche était d'identifier quel modèle peut être considéré comme le meilleur pour résoudre ce problème.
- Pour faire cette déclaration, nous devons vraiment comparer tous les modèles formés en ce qui concerne: la valeur AUC.

L'ÉVALUATION

```
In [252]: # Affichages des résultats par modèles
base_models = models(X,y)
base_models
```

```
Out[252]: (
           Models  Testing Score  Training Score
4      Random Forest           0.951         0.99375
3      Decision Tree           0.924         1.00000
0  Logistic Regression           0.868         0.86425
1              SVM            0.861         1.00000
2              KNN            0.850         0.86750,
None,
           Models  Testing AUC  Training AUC
4      Random Forest    0.847872    0.977993
3      Decision Tree    0.847275    1.000000
0  Logistic Regression    0.603606    0.579305
2              KNN       0.508694    0.553285
1              SVM       0.500000    1.000000)
```

L'ÉVALUATION

Scores de précision par modèle :

Modèle	Score d'entrainement	Score de test
Random Forest	0.99375	0.951
Decision Tree	1.00000	0.924
Logistic Regression	0.86425	0.868
SVM	1.00000	0.861
KNN	0.86750	0.850

Scores de AUC : Aire sous la courbe par modèle

Modèle	Score d'entrainement	Score de test
Random Forest	0.977993	0.847872
Decision Tree	1.00000	0.847275
Logistic Regression	0.579305	0.603606
SVM	0.553285	0.508694
KNN	1.000000	0.500000

L'ÉVALUATION

- Selon les tableaux précédant, les deux méthodes (Random Forest & Decision Tree) montrent une exactitude relativement bonne et similaire pour les données de d'entraînement et de test.

CONCLUSION ET PERSPECTIVES

CONCLUSION ET PERSPECTIVES

- L'importance de la prédiction du churn pour le marché des télécommunications ne cesse de croître. La collecte de données devient une tâche quotidienne pour toutes les entreprises, et la valeur de ces données peut provenir de sources multiples.
- La prédiction du churn est en train de devenir l'une de ces sources qui génèrent des revenus pour l'entreprise et d'être capable de prévenir quand les clients vont cesser leur contrat avec la société ouvre la possibilité de renégocier ce contrat afin de fidéliser le client.

STRATÉGIES D'INTERVENTION

- **À court terme :**
- **Récompenser** ces clients pour leur fidélité et atténuer leur insatisfaction grâce à l'utilisation **d'un cadeau automatisé**.
- Des stratégies de **tarification** pourraient également être utilisées :
- Par exemple, une fois que les clients ont effectué plus de 4 appels de service client pendant la durée de leur contrat, **un rabais pourrait être offert** sur leur facture suivante pour dissiper l'insatisfaction du client et faire preuve d'empathie et de reconnaissance de l'expérience client.
- Un e-mail personnalisé et convivial pourrait également être envoyé au client, reconnaissant les problèmes que le client a eu tout en lui notifiant **une récompense et / ou une réduction** sur sa prochaine facture.

STRATÉGIES D'INTERVENTION

- À long terme :
- **Le traitement du langage naturel** pourrait être utilisé pour effectuer une analyse thématique des thèmes les plus communs identifiés dans les appels de service à la clientèle (ou des questionnaires traditionnels pourraient également être utilisés si nécessaire).
- Des stratégies axées sur le produit et le client à long terme pourraient alors être conçues pour réduire le nombre d'appels de service à la clientèle vécus par les clients.

MERCI POUR VOTRE ATTENTION