

PROCESSUS DATA

Jérôme Lacaille

Expert émérite

DataLab Safran Aircraft Engines

Formation DataClimber

2023



- 1. MÉTHODOLOGIE CRISP-DM**
- 2. LES JALONS**
- 3. ORGANISATION DES CHANTIERS D'ANALYSE**
- 4. CONCLUSION**

1

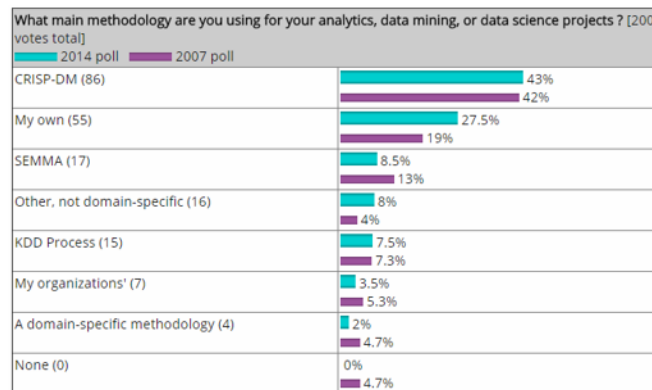
MÉTHODOLOGIE CRISP-DM

Qu'est-ce que CRISP-DM ?

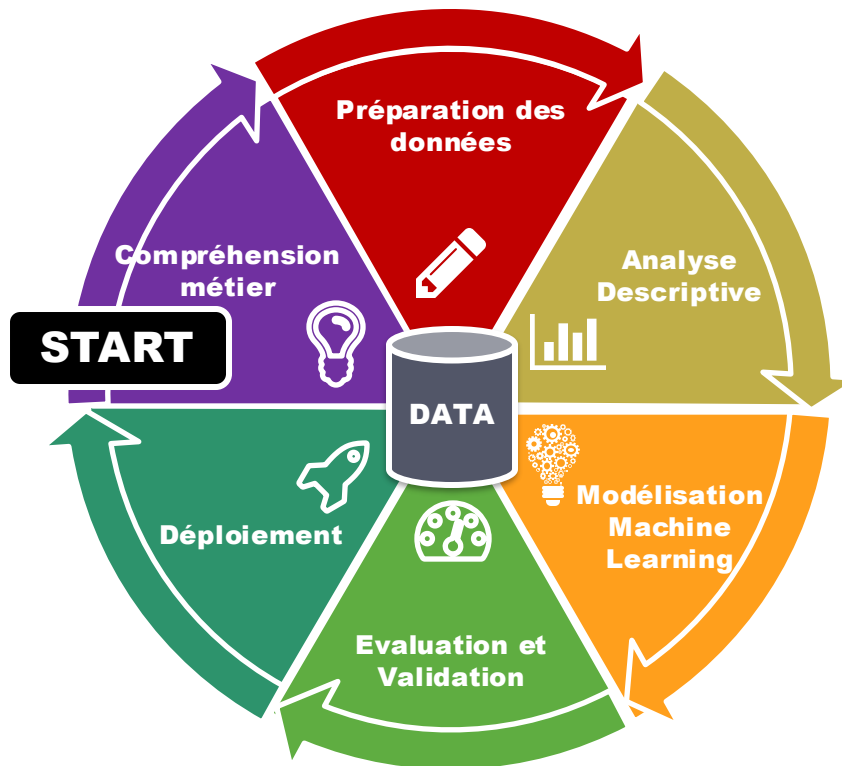
- « **CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter vos travaux d'exploration de données.** » d'après IBM ([lien](#)) :
 - > exploration de données : on cherche à créer de la valeur à partir des données
 - > orienter les travaux : c'est une méthodologie
 - > exploration de données : définition itérative du besoin avec quelques fois des besoins initiaux non clairement identifiés

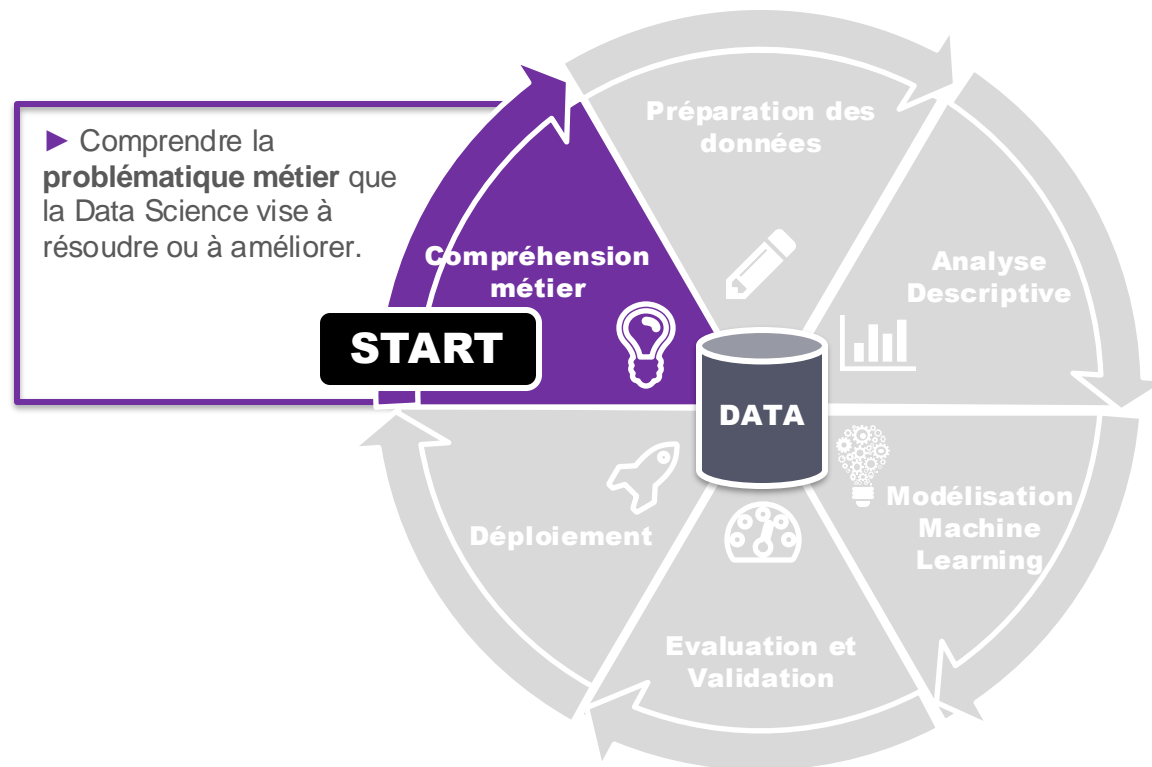
- **CRISP-DM est une méthodologie utilisée :**

<https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>

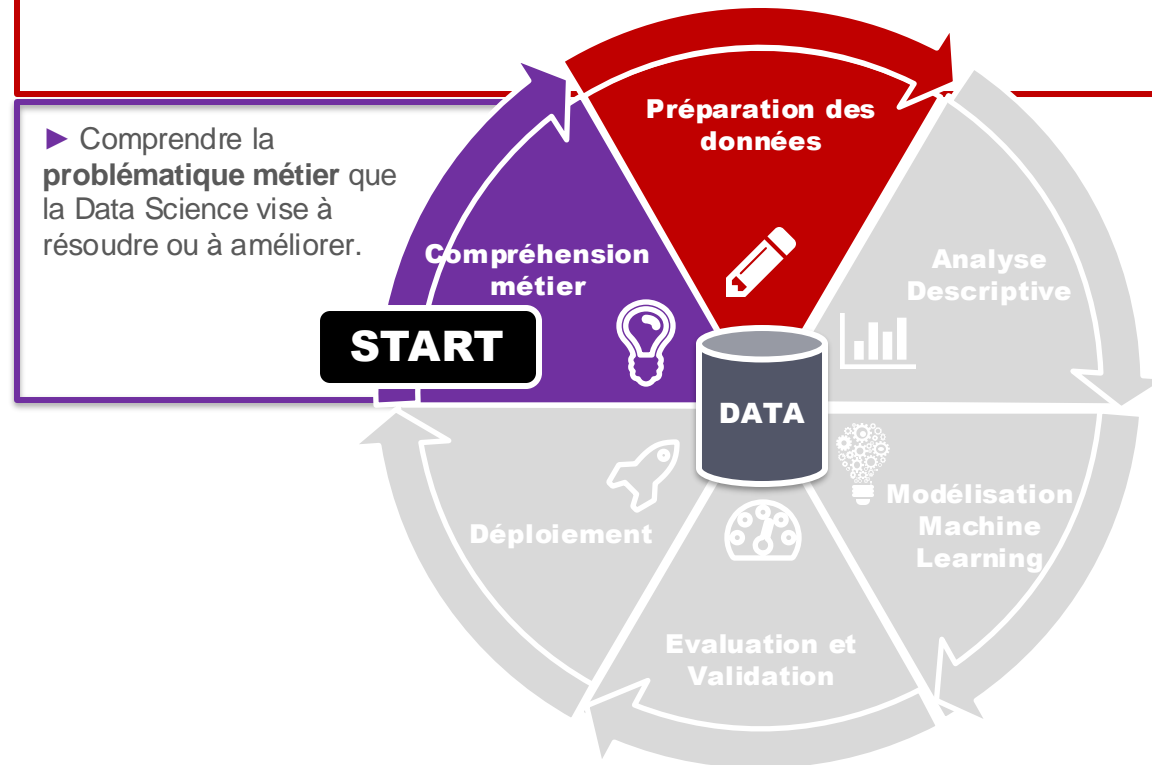


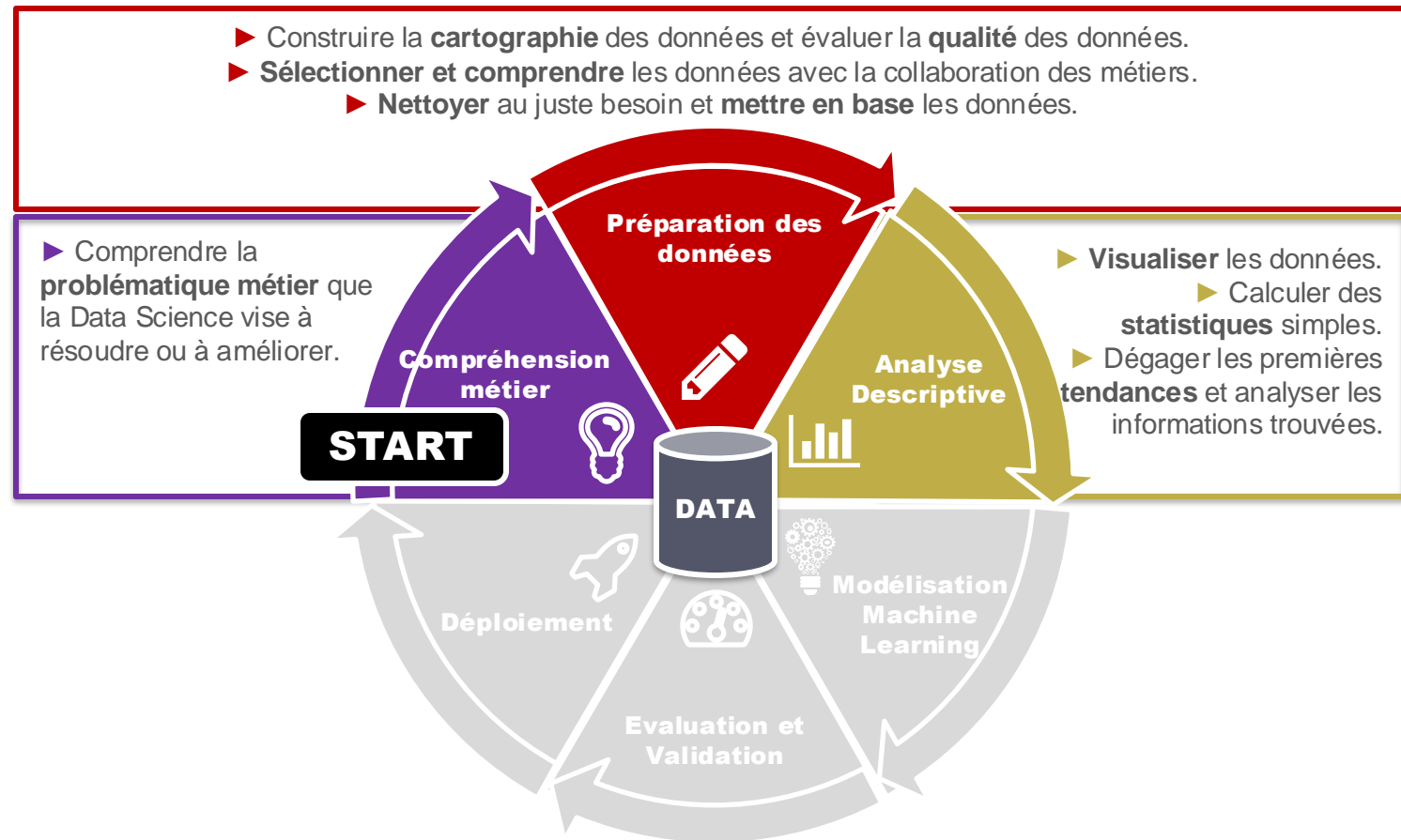
Le processus CRISP-DM

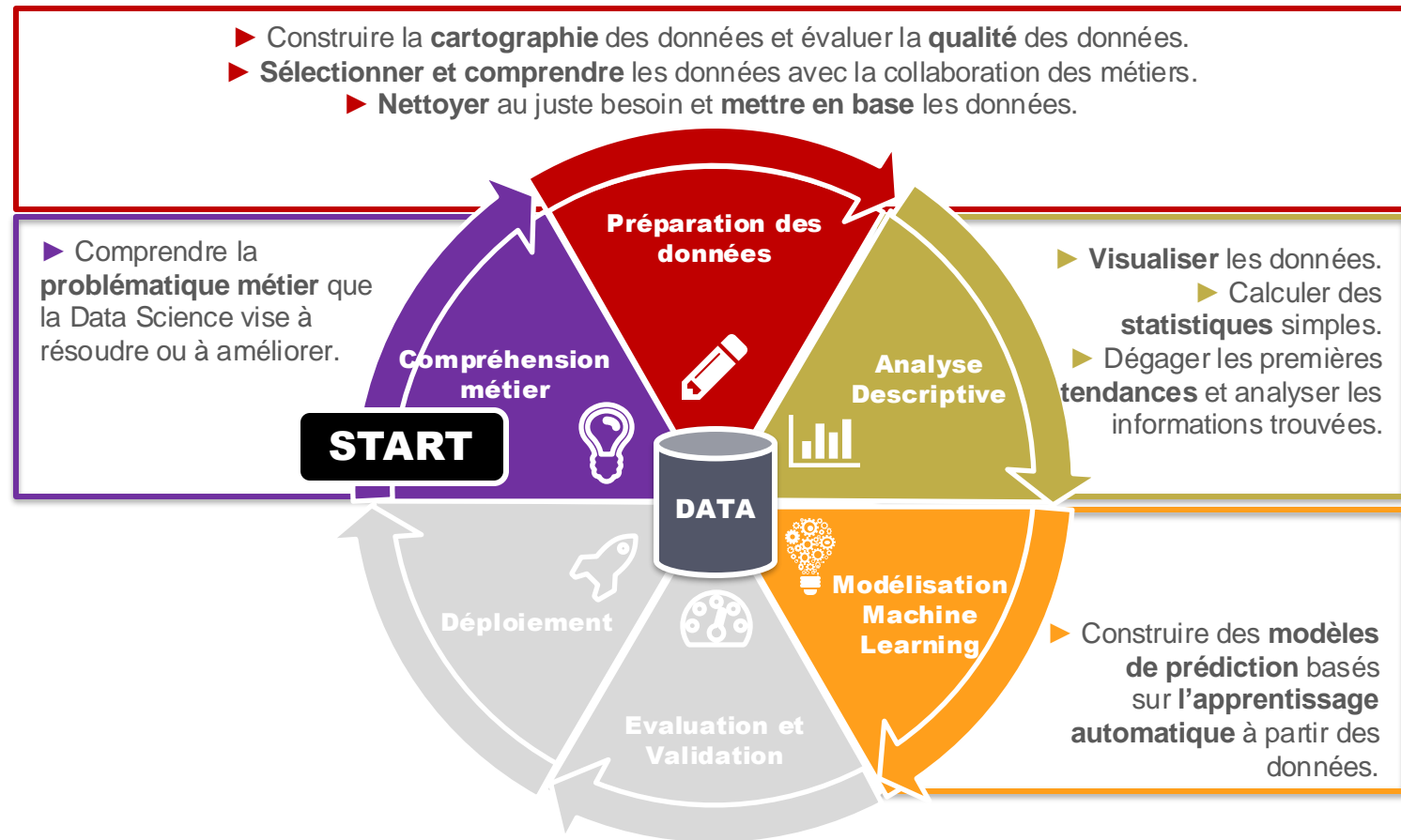




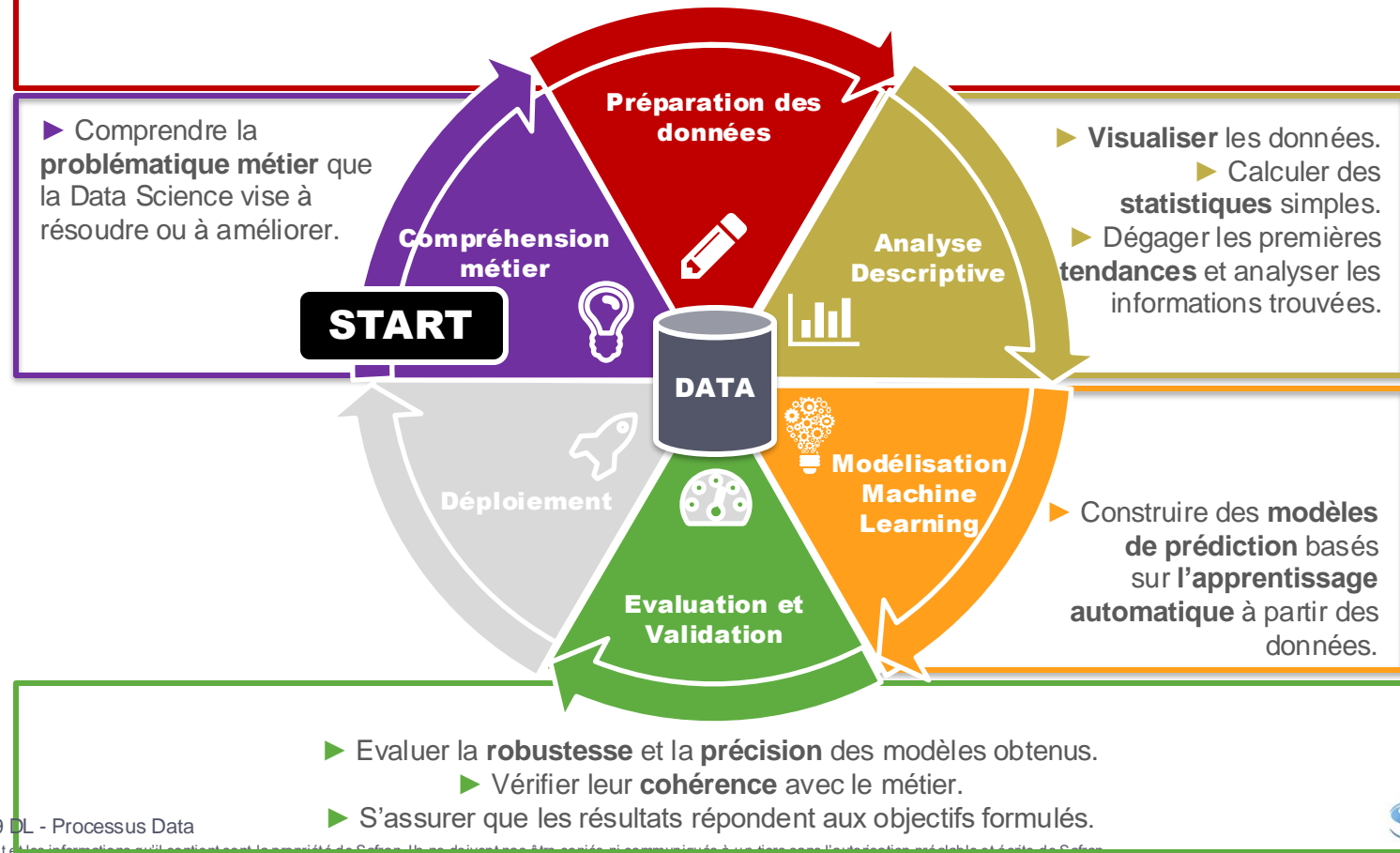
- ▶ Construire la **cartographie** des données et évaluer la **qualité** des données.
- ▶ **Sélectionner et comprendre** les données avec la collaboration des métiers.
- ▶ **Nettoyer** au juste besoin et **mettre en base** les données.



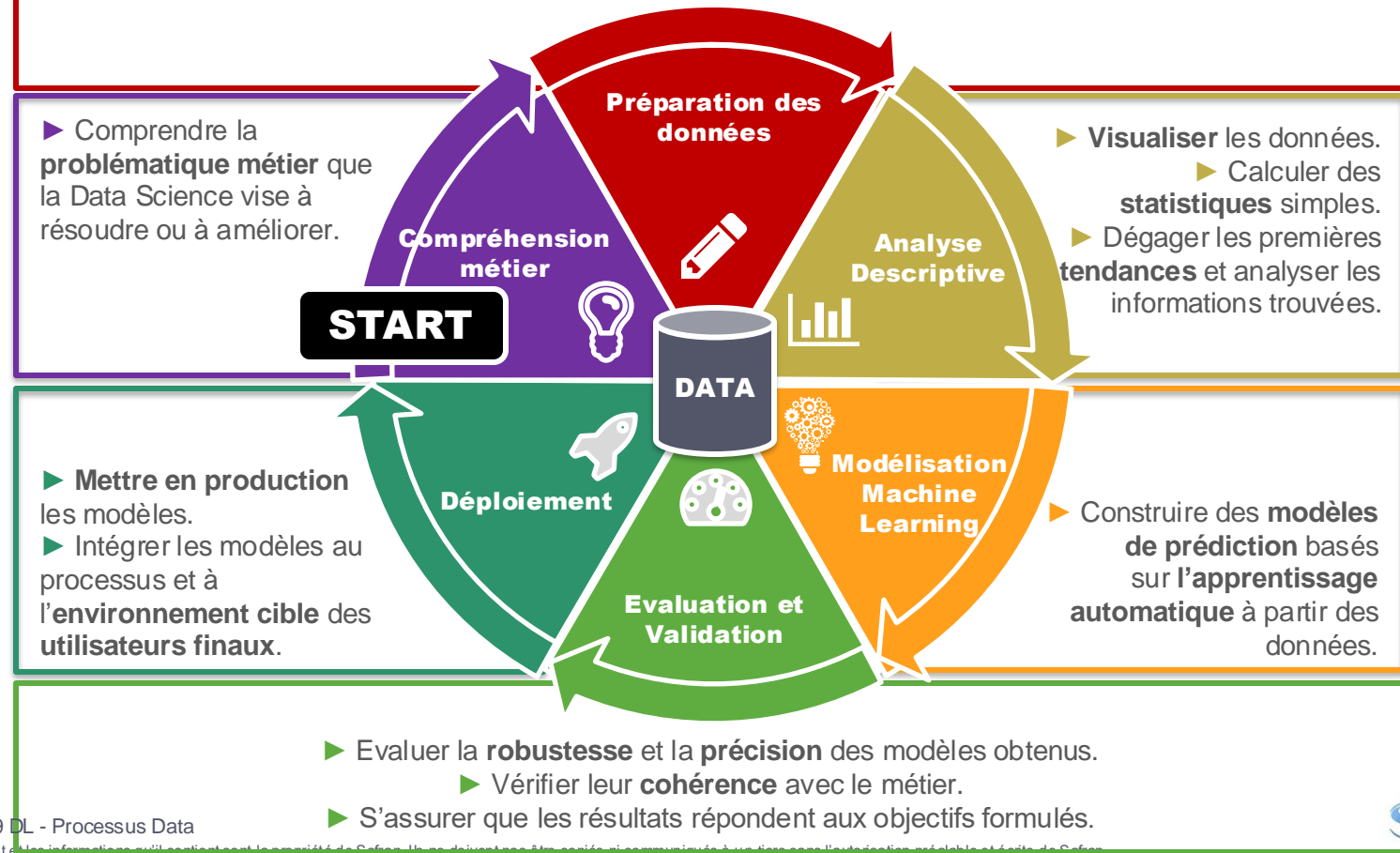




- ▶ Construire la **cartographie** des données et évaluer la **qualité** des données.
- ▶ **Sélectionner et comprendre** les données avec la collaboration des métiers.
- ▶ **Nettoyer** au juste besoin et **mettre en base** les données.



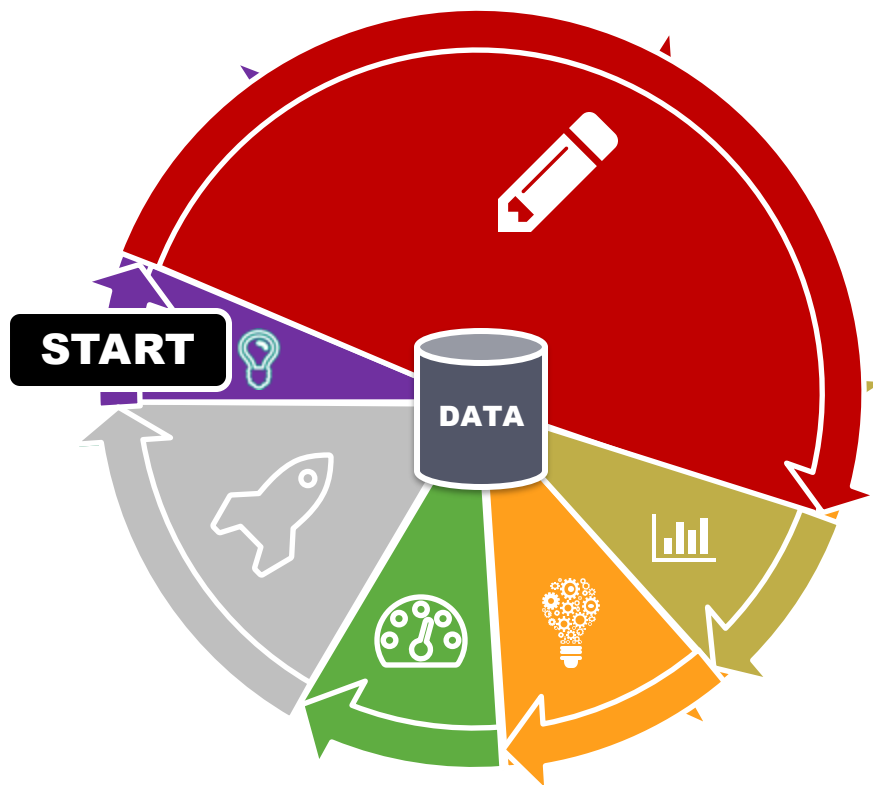
- ▶ Construire la **cartographie** des données et évaluer la **qualité** des données.
- ▶ **Sélectionner et comprendre** les données avec la collaboration des métiers.
- ▶ **Nettoyer** au juste besoin et **mettre en base** les données.





Processus CRISP-DM

Cross Industry Standard Process for Data Mining



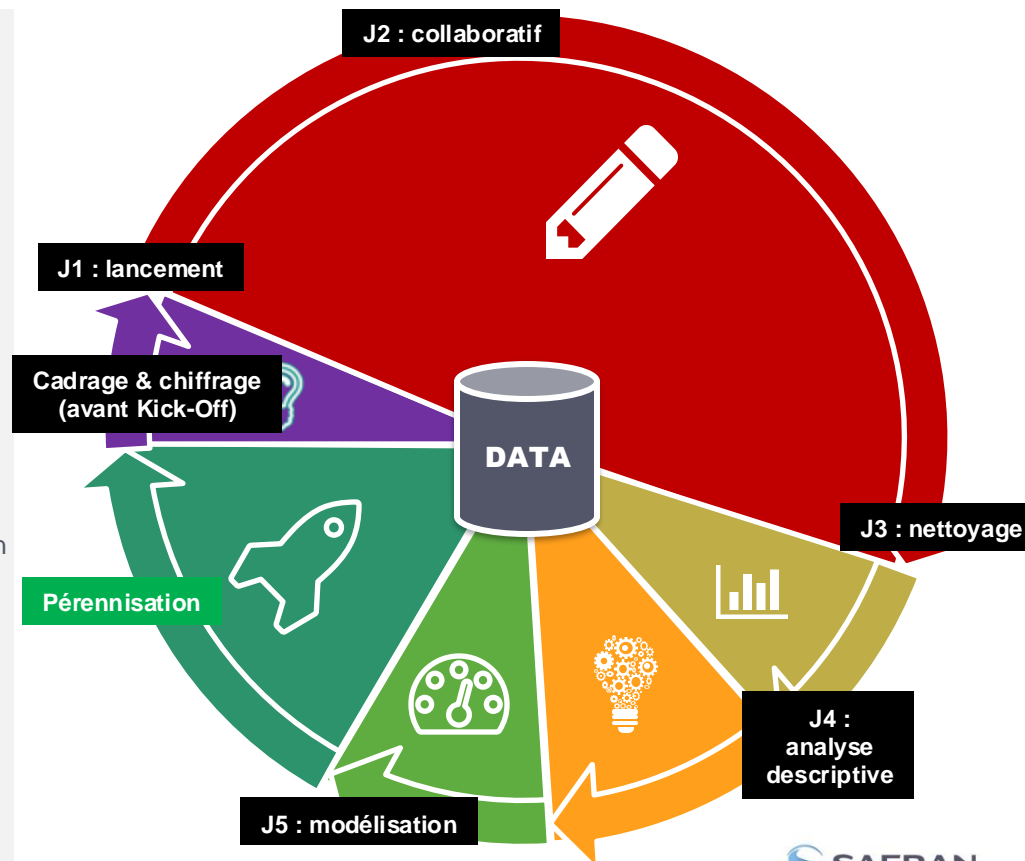
Processus CRISP et Jalons

5 jalons à l'analyse de données

1. Jalon « **lancement** »
Enoncé du problème, du plan de travail, cartographie des données, organisation de l'équipe, planning, analyse de risque projet
2. Jalon « **collaboratif** »
Plan de collecte des données et son exécution, définition « métier » d'indicateurs, capitalisation
3. Jalon « **nettoyage** »
Mise en base des données suivant une série d'opérations (extraction, parsing, mise en base) et évaluation de la qualité de la donnée ainsi nettoyée
4. Jalon « **analyse descriptive** »
Compréhension (statistique) les données, des dépendances entre variables. Réponse au problème ou préparation pour la modélisation
5. Jalon « **modélisation** »
Livraison d'un modèle, justification des choix de conception, validation et présentation des résultats vs problème

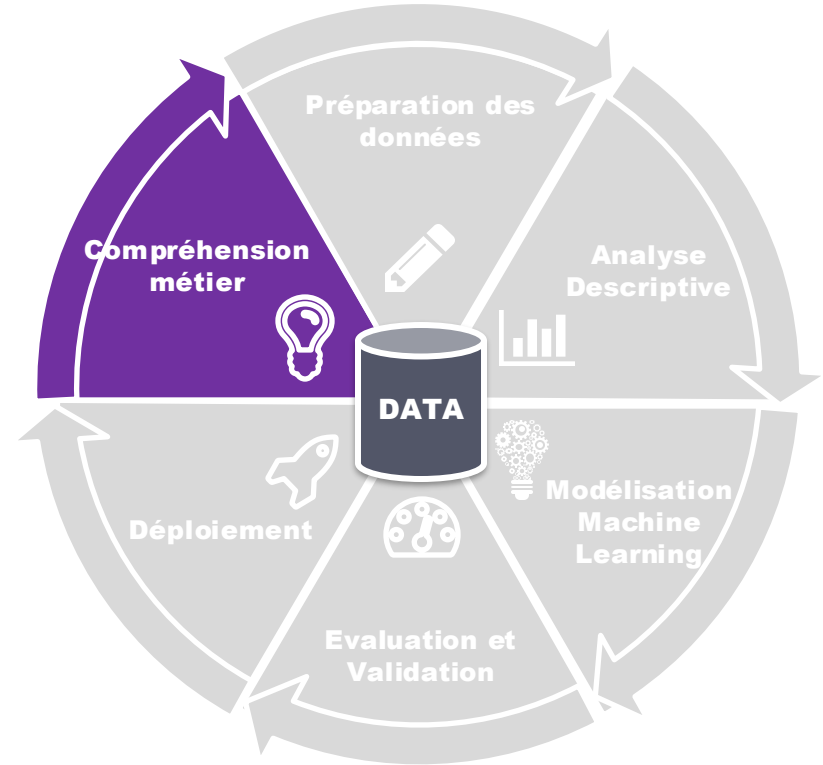
> + 1 Jalon « **Pérennisation** »

- > Rédaction d'un résumé du chantier, présentation des résultats du chantier à son secteur, choix de la voie à suivre après le chantier



2.1

J1 : KICK OFF



Énoncé du problème

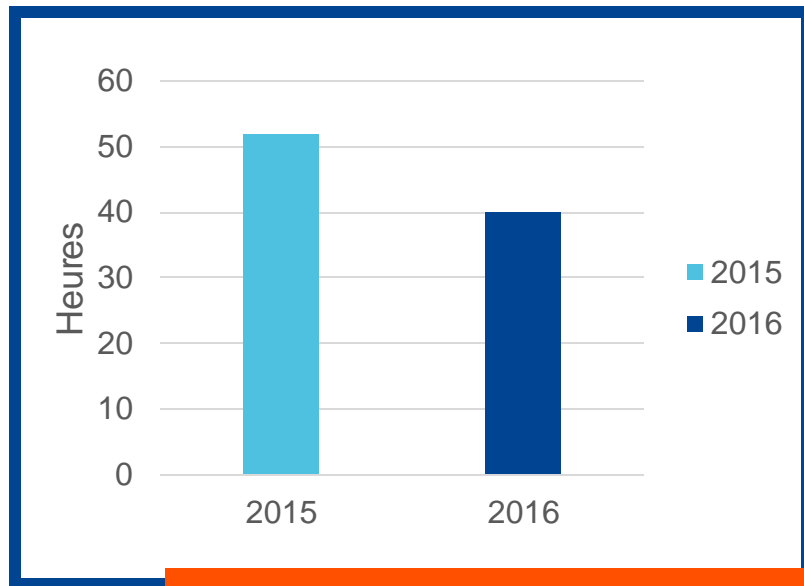
▪ Situation actuelle

- D'avril 2015 à avril 2016
 - ♦ En moyenne XX % des temps d'assemblage dépassent 52 heures
 - ♦ Le standard était estimé à 40 heures
- Les données sont issues de SAP qui enregistre les déclarations quotidiennes du personnel

▪ Principales conséquences

- La capacité de production ne permet pas de suivre la demande
 - ♦ 600 modules prévus en 2015
 - ♦ seulement 480 modules livrés l'an dernier
- Notre ligne d'assemblage perd de la compétitivité vis-à-vis d'autres centres de production
- La Perte financière se monte à 500 k€

Objectifs et impact du projet



▪ Objectifs du projet

- Réduire en 4 mois la durée moyenne d'assemblage des modules de 52 à 40 heures

▪ Impact du projet et gains estimés

- Amélioration du confort de travail pour les compagnons par la réduction des urgences
- Accélération du flux de modules produits, augmentation de l'On Time Delivery et satisfaction de la demande clients
- Retour à l'équilibre financier (+ 500 k€)

Même s'il est encore difficile d'estimer les gains d'un projet d'analyse de données, le faire au début du projet est très important. Cela permettra de valoriser son projet.

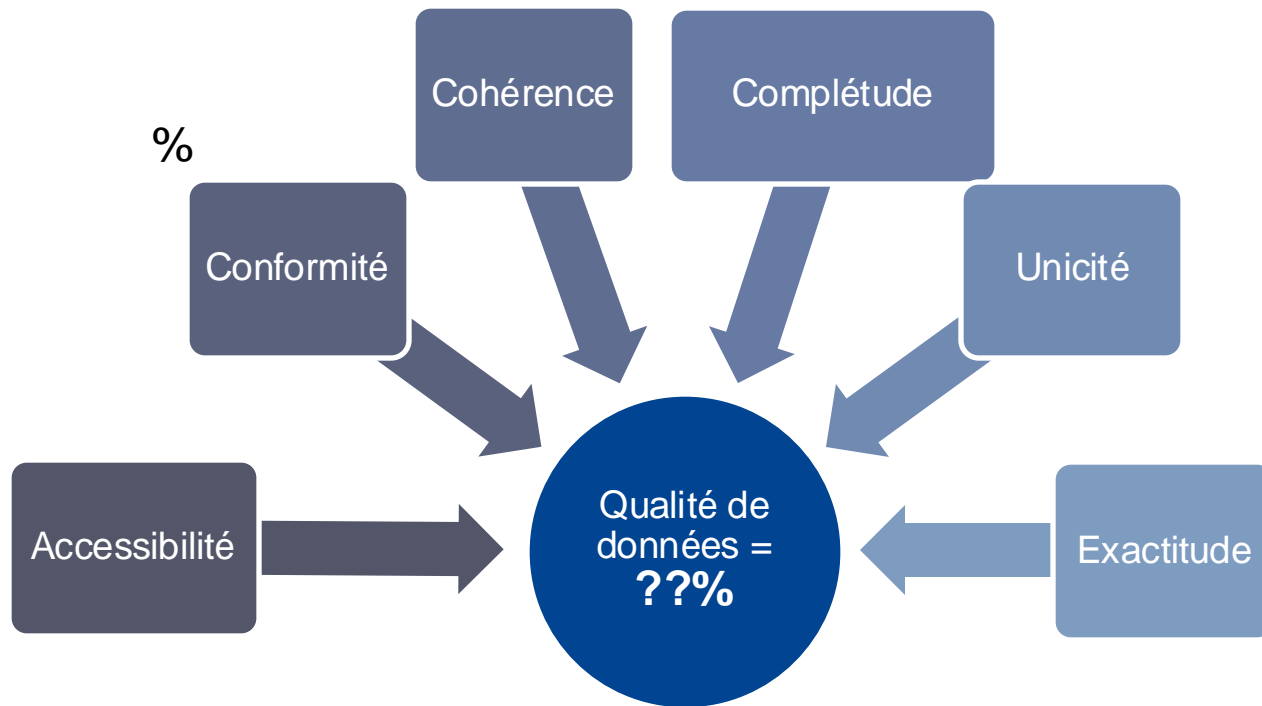
Cartographie des données

Fournisseurs	Dénomination	Description	Emplacement (Data Sore, SGBD,...)	Quantité / Volume (nb variables, nb observations, ...)
YT	Point stabilisés	Essais perfo SCR 24-3	Moise base thermo	120 param /420Mo

La démarche de collecte des données est portée collectivement par l'équipe projet avec les métiers.

La couverture doit d'abord être large (cartographie des données) puis être spécifique (sélection des données dans le jalon collaboratif).

Evaluation de la qualité des données d'entrée



- La charge de travail attendue à l'étape de préparation des données est très liée à la qualité des données brutes.
- Il est recommandé d'évaluer la qualité des données brutes le plus tôt possible avec un échantillon représentatif.
- Pour rappel, à ce stade là, les données brutes ne sont pas encore extraites d'où la nécessité d'utiliser un échantillon

Planning du projet

■ Prévission initiale : déroulement sur XX semaines



■ Hypothèses de planification

Jalon	Charge	Commentaire
Lancement	1h	
Collaboratif	2 à 3h	Avec l'équipe projet
Nettoyage	60h	Dont 2h avec équipe projet
Analyse	40h	Dont 3h avec équipe projet
Modélisation	10h	Dont 1h avec équipe projet

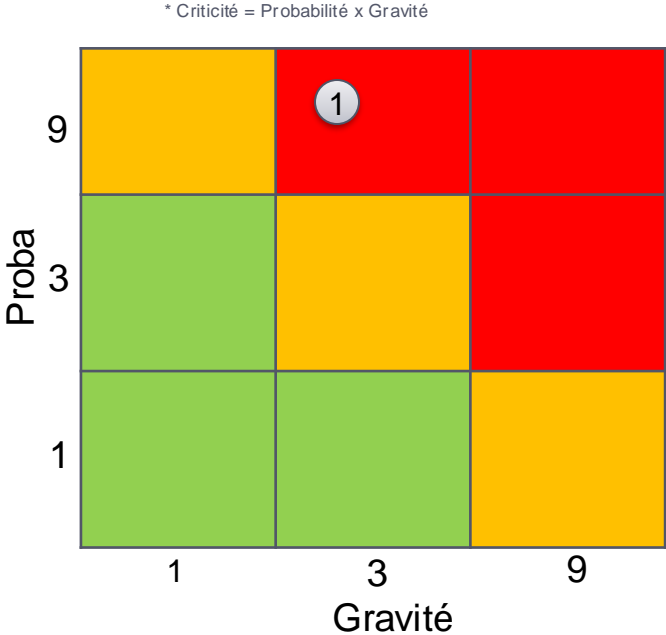
- Attention à ne pas sous-estimer la charge de travail attendu. Une fois que vous avez défini le plan de travail, faite un chiffrage.
- D'une façon générale, l'étape de Nettoyage est l'étape la plus longue.

Equipe projet, rôles et responsabilités

Rôle dans le projet	Prénom	Nom	Département	Fonction
Chef de Projet	Xx	Yy		Data scientist
Tuteur Data scientist	Xx	Yy		Data scientist
Implémentation du Processus de Changement	Xx	Yy		L928 Support Utilisateur
Customer communication envers le client	Xx	Yy		L928 Interface HMG
Processus Global	Xx	Yy		L928 Design
Analyse Qualité	Xx	Yy		A clarifier
Analyse de l'habillage	Xx	Yy		L928 Design
Analyse de la Préparation	Xx	Yy		L928 Design

Analyse des risques projet







Num éro	Risques projet	Plan de levée des risques
1	Lorem ipsum dolor sit amet...	Lorem ipsum dolor sit amet...



Organisation du projet

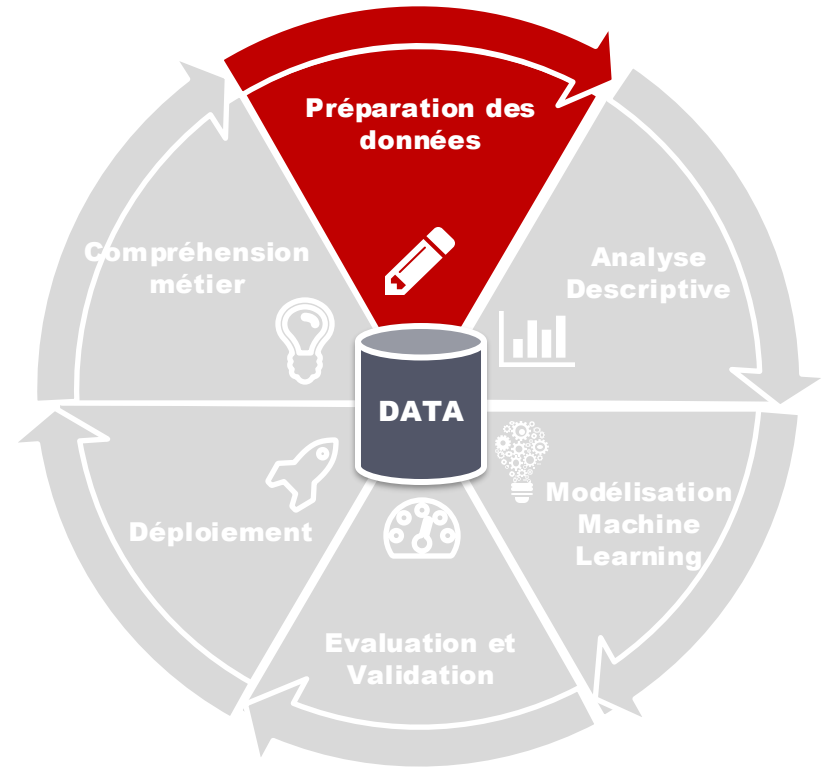
Quoi	A Qui	Quand	Resp.	Comment	Où	Commentaires
Réunions d'équipe	Equipe	Hebdomadaire	Jeff	Réunion	Tableau	
Compte rendu de réunion	Equipe	Hebdomadaire	Jeff	Email		
Liste d'Actions	Equipe	Hebdomadaire	Jeff	Email		
Revue de Projet	Managers	Mensuel	Jeff	Présentation		
Rapport d'avancement	Sponsor	Bihebdomadaire	Jeff	Réunion	Bureau du Sponsor	

Franchissement du jalon

Critère	Réalisé
Jalons planifiés, problème et objectif complétés et approuvés	
Cartographie des données	
Evaluation de la qualité des données	
Equipe projet approuvée et opérationnelle	
Risques majeurs et plan de maîtrise associé	
Organisation du projet	

2.2

J2 : COLLABORATIF



BRAINSTORMING : paramètres à collecter

Variables de sortie

(Qu'est-ce que je veux expliquer, prédire, classifier ?)

Nom de la variable	description

Variables d'entrée

(Qu'est-ce qui peut jouer sur mes sortie ?)

Nom de la variable	description	Influe sur la sortie a priori (O/N)
N1(t)		
N1 max par cycle		
N1 moyen par heure		

S'appuyer sur la cartographie des données !!

Paramètres de processus

(Quels sont les réglages, moyen de mesure, ... qui peuvent jouer sur ma sortie ?)

Nom de la variable	description	Influe sur la sortie a priori (O/N)

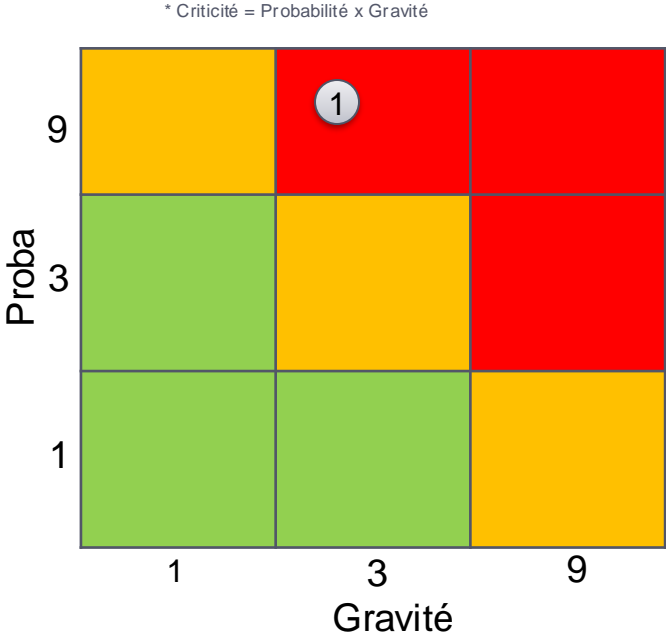
Plan de collecte / plan d'expérience

Indicateur	Base de donnée source	Nom du responsable
Données géométriques sélectionnées	Caracas	Robert
Données d'essais de réception sélectionnées	Moise	Michel




Au jalon collaboratif, le plan de collecte doit être finalisé à 100%.

MaJ de l'analyse de risque

Num éro	Risques projet	Plan de levée des risques
1	Lorem ipsum dolor sit amet...	Lorem ipsum dolor sit amet...



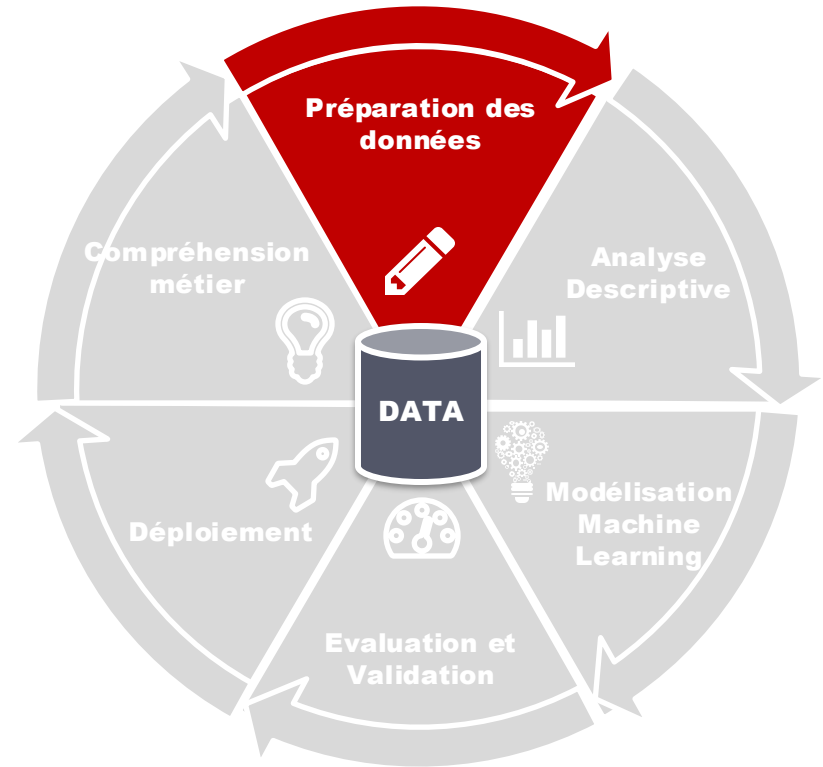
Le principal risque à cette étape est de ne pas pouvoir comprendre et maîtriser l'ensemble des données et donc à fortiori ne pas savoir sélectionner les paramètres pertinents. Bien évidemment, un tel risque pourra remettre en cause la poursuite du projet.

Critère	réalisé
Le plan de collecte a été réalisé avec l'ensemble des parties prenantes	
Les données qui seront utilisées sont décrites dans un dictionnaire	
La définition des indicateurs est claire et partagée avec l'ensemble de l'équipe projet	

- **Bien penser à la suite du jalon collaboratif d'informer le data steward (ou à défaut le data manager) d'un possible ajout ou modification du dictionnaire des données.**

2.3

J3 : NETTOYAGE DE DONNÉES



1- Extraire les données

Reference de la base prélevée (ou Référence du fichier)	Date de l'extraction (ou Date de fourniture du fichier)	Responsable de la donnée fournie	Niveau de validation connu (validé SAE/ validé Y/ validé partenaire/ pas de validation connue)	Liste des identifiant moteurs, serial number, part number, machine,...	Dates

- Extraction effective de la donnée : tracer la méthodologie de l'extraction et le niveau de validation
- Analyser des données avec un niveau de validation inconnu est très risqué. Dans ce cas, il est légitime de remettre en question la poursuite du projet.

Parsing / croisement des données / calcul d'indicateurs

- **Utiliser et mettre les codes sous GIT / adresse**
- **Décrire la méthode utilisée (code utilisé, calcul des indicateurs)**
 - Définition des indicateurs
 - ...

Nettoyer les données

1. Eliminer les données inutiles

1. Ligne de zéro?
2. Doublons

2. Classer les données par niveau de crédibilité

1. Données de bonne qualité
2. Données de qualité moyenne
3. Données suspectes

3. Identifier les données manquantes

1. Ex : extrapolation, calculs intermédiaires, compléments d'information

Présenter les méthodes de calcul intermédiaires au jalon

Mettre en base

- **STANDARD :**
- **Choix de la base selon une checklist de possibilités techniques**
 - MongoDB ?
 - MySQL ?
 - Fichier Excel ?
 - ...
- **Comment accéder à la base de données nettoyées (lien Serveur,...)**

Evaluer la qualité des données de sortie



☐ Reprendre l'évaluation de la qualité après nettoyage

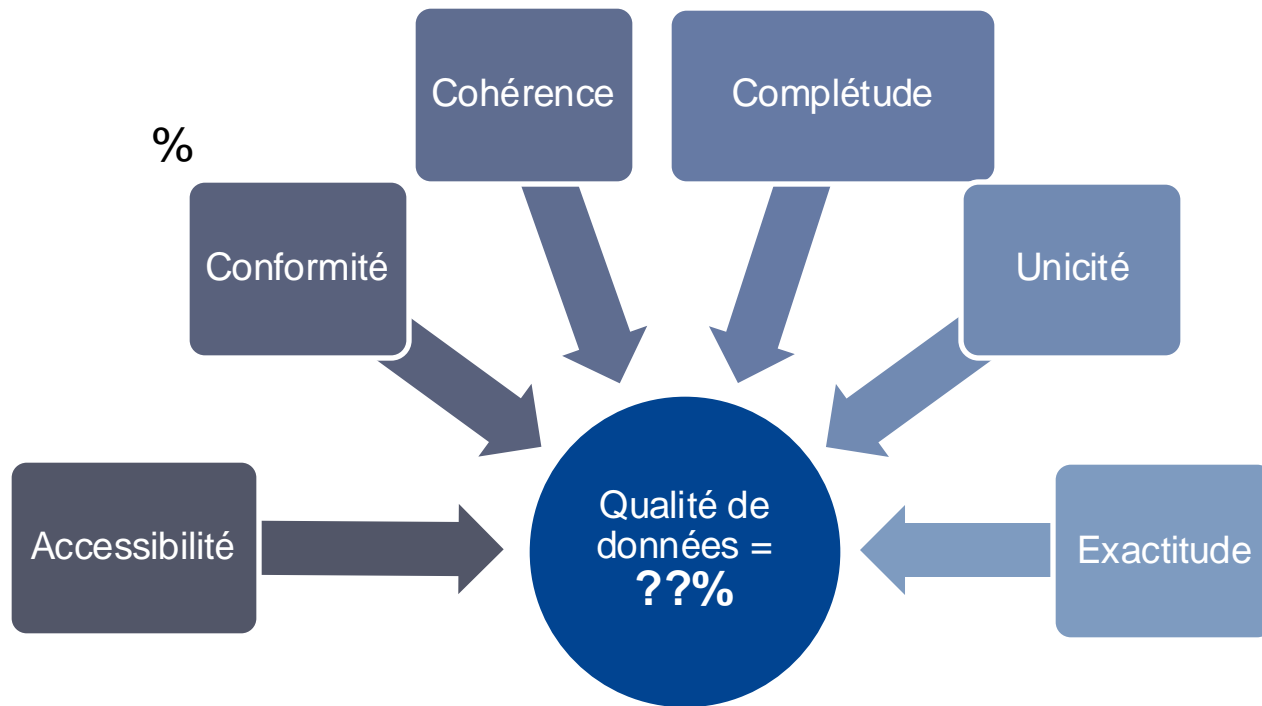
Qualité obtenue	Description de la qualité	Niveau
Xx %	Données en base, de sources fiables, valeurs aberrantes identifiées,...	

☐ Action d'amélioration de la qualité :

- ☐ Exemple : demander une validation de la base de données d'entrée, prévenir le client, etc...

- La qualité des données après les étapes de nettoyage doit atteindre au moins le seuil de 80%.
- Il est très important de capitaliser les actions de nettoyage réalisées et les axes d'amélioration encore possibles pour partager avec les acteurs de la data gouvernance.

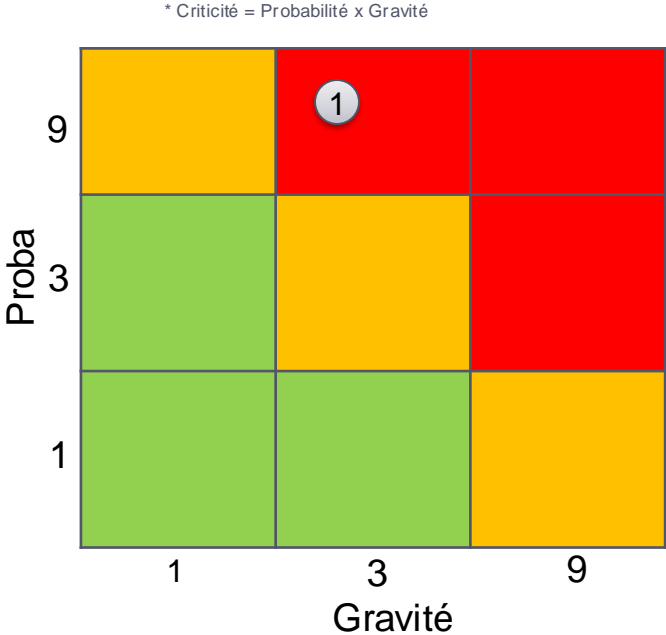
Evaluation de la qualité des données après nettoyage






Produit des % par catégorie

MaJ de l'analyse de risque

Num éro	Risques projet	Plan de levée des risques
1	Lorem ipsum dolor sit amet...	Lorem ipsum dolor sit amet...

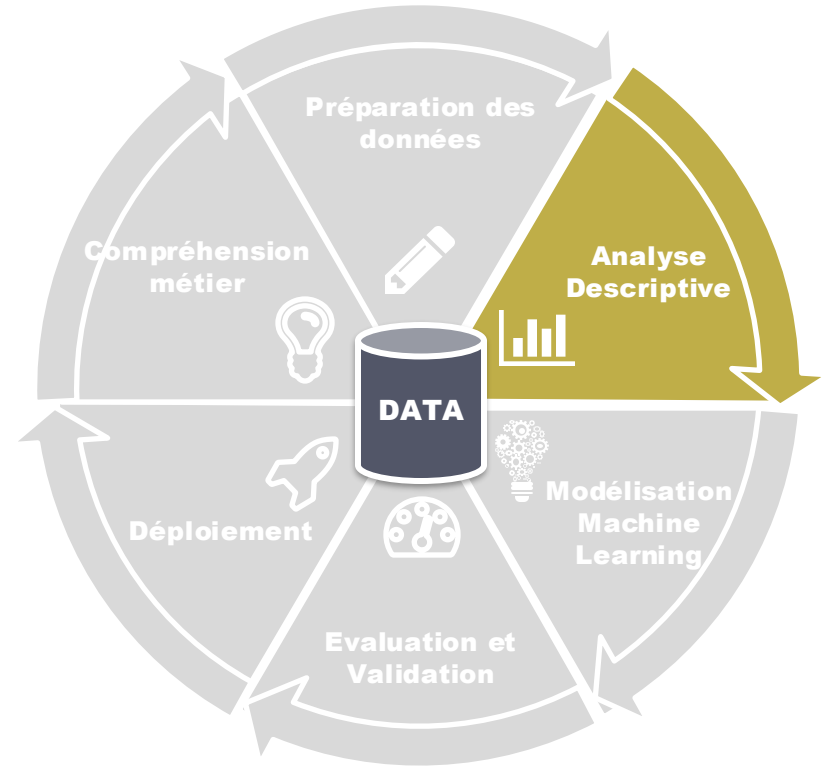


Critère	réalisé
La qualité de données après nettoyage est au dessus de 80%	
Les règles de nettoyage ont été décrites et le code associé est sous git. Le Data Steward (ou à défaut le Data Manager) a été informé	
Les non-qualités et les actions d'amélioration ont été remontées auprès du Data Steward (ou à défaut le Data Manager)	

2.4

JALON

ANALYSE DESCRIPTIVE



Objectifs et démarche

- **Comprendre et caractériser les variables**

- Distribution des variables,
- Evolution des variables (série temporelle),
- Moyenne, dispersion, ...
- Identifier les facteurs discriminants,
- Corrélation entre les variables,
- ...

L'analyse descriptive permet de comprendre les données et les interactions entre celles-ci. De fait, elle est incontournable dans l'étude.

Description des données en entrée

- **Population (contexte)**
 - Exemple : date d'essai, date de vols, compagnie,
- **Variables étudiées**
 - Nature des variables : quantitative, qualitative
 - Définition de la variable (unité ...)
 - Logique d'enregistrement
- **Echantillon disponible (taille, obtention)**

Vérifier la couverture des données disponibles et évaluer la faisabilité du projet.

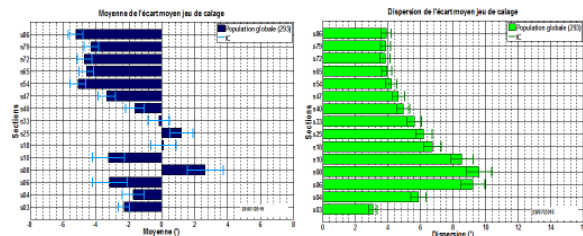
Résultats d'analyse

Résultats

- Distribution des variables,
- Evolution des variables (série temporelle),
- Moyenne, dispersion, ...
- Identifier les facteurs discriminants,
- Corrélation entre les variables,
- ...

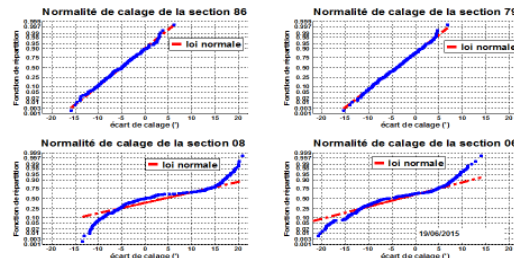
Il est possible de prévoir la fin du projet au jalon descriptif. Dans ce cas-là, il faudra s'assurer, dans le plan de travail, que l'on réponde à l'objectif.

Δ_CALAGE – POPULATION GLOBALE PÉRIODE, TOUS USINEURS
MOYENNE ET DISPERSION



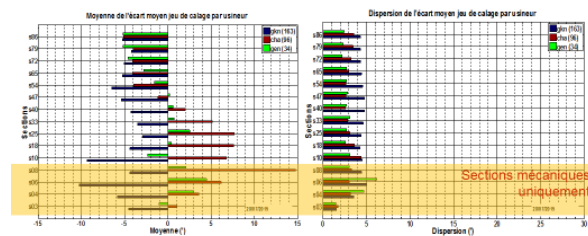
- **Moyenne**
 - Sections hautes : moyennes homogènes sur la hauteur
 - Sections basses : dispersées sur la hauteur
- **Dispersion**
 - Sections basses : fortement dispersées, sauf la section 03

Δ_CALAGE – POPULATION GLOBALE
TEST DE NORMALITÉ DE LA DISTRIBUTION (1)



- Sections hautes : Distributions des écarts de calage proches de la loi normale.
- Sections basses : Distribution loin d'une loi normale → Il y a donc vraisemblablement des biais de procédé en sections basses.

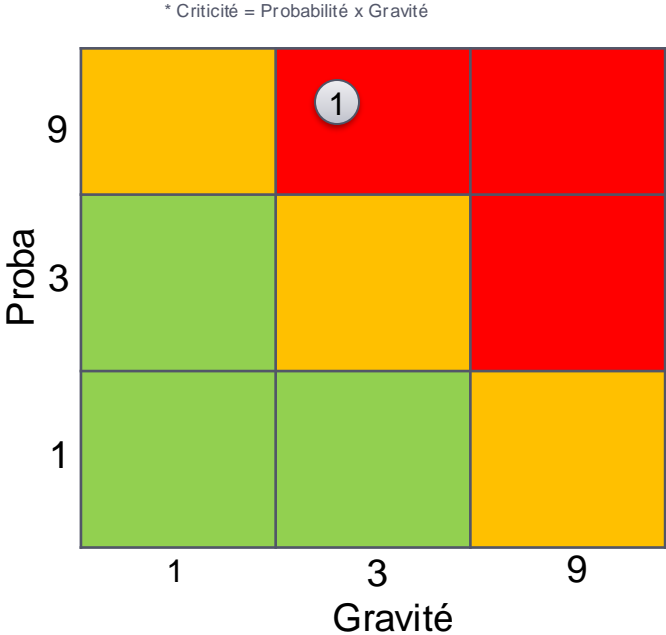
Δ_CALAGE – POPULATION GLOBALE PÉRIODE, PAR USINEUR
MOYENNE ET DISPERSION






- **Moyenne**
 - Sections hautes : 3 fournisseurs identiques
 - Sections moyennes et basses : Fort écart (>3 écart-type).
 - CHA plus élevée – GKN plus faible
- **Dispersion**
 - Dispersion beaucoup plus faible en section basse.
 - Peu variable sur la hauteur (sauf pied)
 - GKN plus dispersée

MaJ de l'analyse de risque

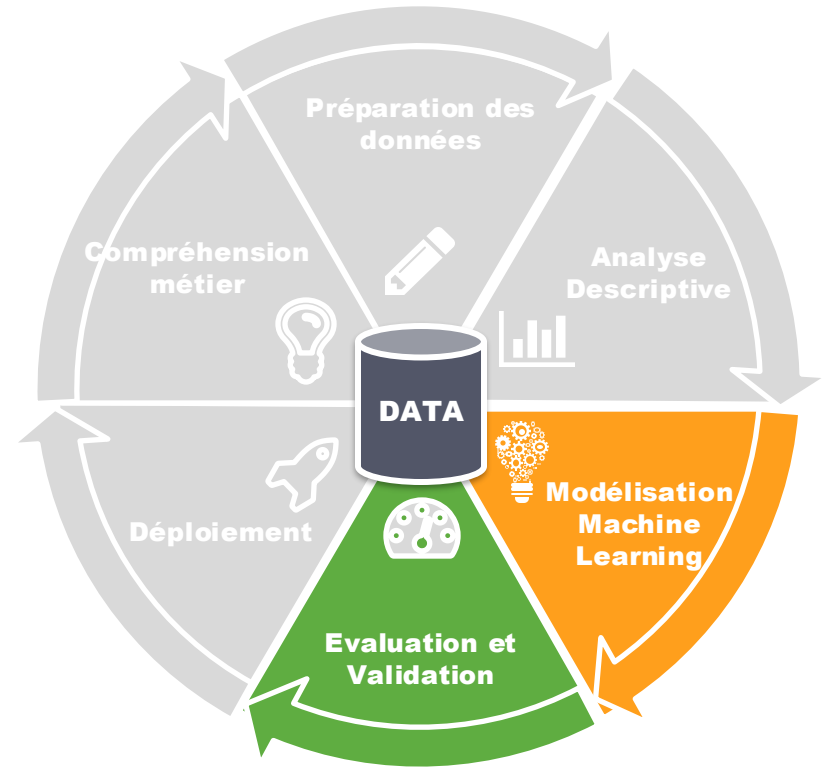
Num éro	Risques projet	Plan de levée des risques
1	Lorem ipsum dolor sit amet...	Lorem ipsum dolor sit amet...



Critère	réalisé
La stratégie d'analyse descriptive a été expliquée	
Les méthodes d'analyse descriptives ont été appliquées (analyse des corrélations, analyse des distribution, ACP,...)	
Les métiers valident les conclusions de l'analyse	

2.5

J5 : MODÉLISATION



Objectif & démarche

- **Modéliser le lien entre la variable cible et les paramètres influents**
 - Définir la variable cible (continue, discrète, ordinale,...)
 - Sélectionner les variables influentes
 - Identifier les types de modèles adaptés à la problématique
 - Définir les critères de performances et la procédure de validation
 - Résultats

Description du modèle

- **Variable cible / Variables explicatives**

- **Quelle méthode ?**

- Modèle linéaire
- Arbre de décision/régression
- Réseaux neurones

- **Paramètres du modèle**

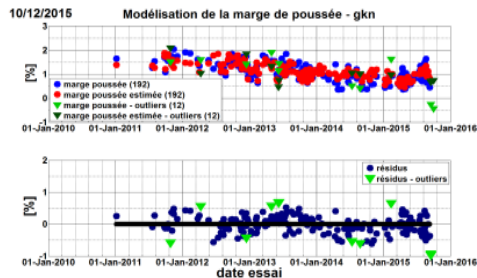
- **Justification du choix de la méthode**

Implémentation et validation du modèle

- **Définir les critères de performance**
 - Exemple : minimiser l'écart type et résidu centré
- **Éliminer les variables suffisamment liés à d'autres pour perturber le modèle**
- **Sélectionner les variables les plus pertinents pour expliquer la variable cible**
- **Estimation des paramètres du modèle**
- **Optimiser le modèle en éliminant les outliers**
- **Tester la robustesse de la prévision du modèle**

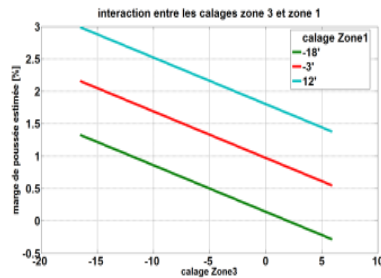
Utilisation opérationnelle du modèle

MODÈLE PARAMÉTRIQUE DE PRÉDICTION DE LA POUSSÉE MODÉLISATION DE L'ÉVOLUTION DE LA MARGE DE POUSSÉE GKN



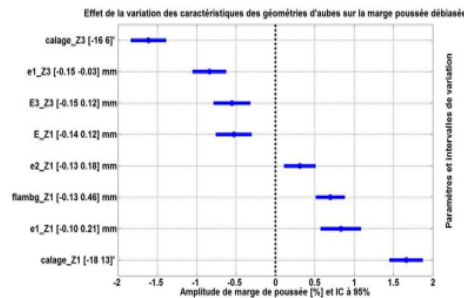
- Le modèle suit la dérive de la marge de poussée
- La marge de poussée est surestimée sur la période 2014
- Les outliers sont répartis sur toute la période des essais

UTILISATION OPÉRATIONNELLE DU MODÈLE DE CORRÉLATION INTERACTION ENTRE LES CALAGES DES ZONES 3 ET 1



- L'effet de variation de calage de la zone 3 sur la marge de poussée est d'une amplitude de ~1,6% quel que soit la valeur du calage zone 1
- Cependant, l'offset de variation de la marge de poussée est différent selon la valeur du calage zone 1

UTILISATION OPÉRATIONNELLE DU MODÈLE DE CORRÉLATION SENSIBILITÉ AUX CARACTÉRISTIQUES GÉOMÉTRIQUES



- Les 2 paramètres les plus sensibles sont le calage en tête et en pied (sensibilité opposée), puis les épaisseurs BA (e1)
- Les autres paramètres montrent une influence réelle mais moins importante

22 / CONFIDENTIEL / JYSS/MMS/GACM



22 / CONFIDENTIEL / JYSS/MMS/GACM

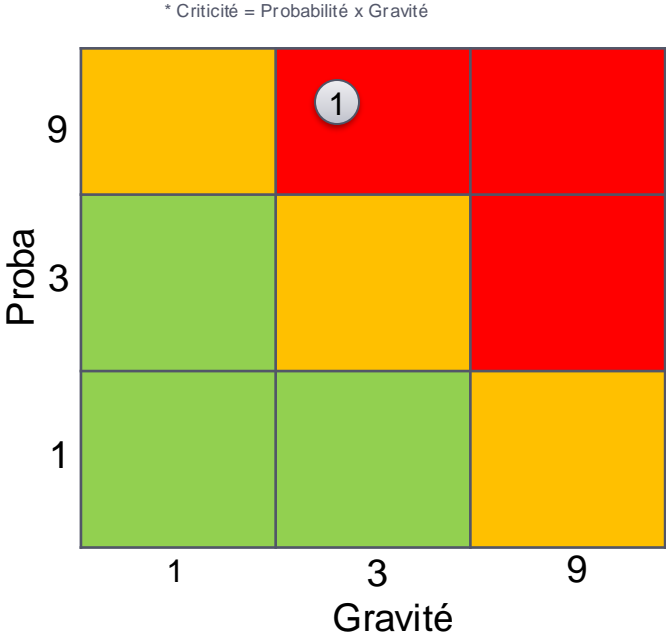


21 / CONFIDENTIEL / JYSS/MMS/GACM







MaJ de l'analyse de risque

Num éro	Risques projet	Plan de levée des risques
1	Lorem ipsum dolor sit amet...	Lorem ipsum dolor sit amet...

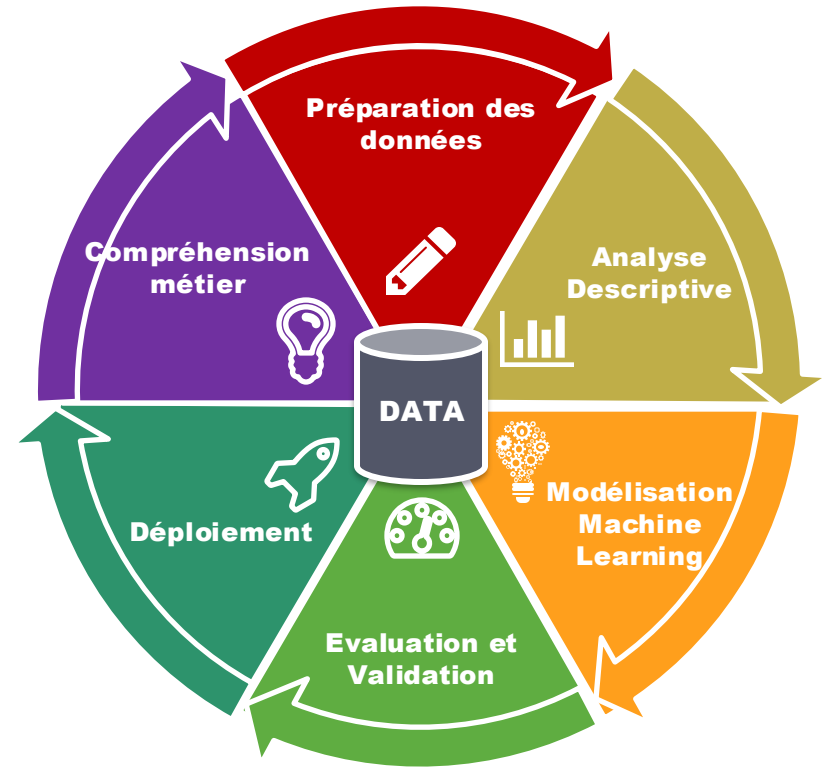


Franchissement du jalon

Critère	réalisé
Le choix du modèle a été expliqué (classification, régression, supervisé, non-supervisé,...)	
Les critères de performance du modèle ont été définis et calculés	
Une stratégie de validation du modèle a été appliquée (validation croisée, ...)	
Les métiers valident le modèle	

2.6

RESTITUTION & PÉRENNISATION



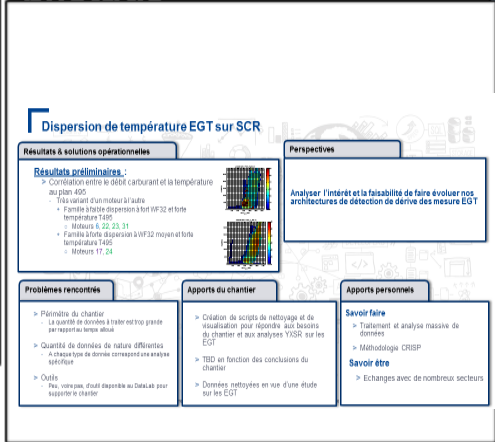
Rédiger un document résumé du chantier

- Problématique et Objectif
- Description des données
- Plan de travail
- Résultats

Choisir la voie à suivre après le chantier



Présenter les résultats du chantier au comité de pilotage



Il s'agit d'une étape pour pérenniser le travail réalisé, décider de la suite à donner à son parcours projet et personnel et faire une restitution de son chantier au comité de pilotage.

2.7

DÉPLOIEMENT (HORS CHANTIER DATA)

La phase de déploiement

A ce stade, nous devons nous assurer que le modèle mis en place est :

- performant (validation de la performance avec le client)
- exploitable = les données existent et sont propres ou nettoyables
- exploitable sur un problème opérationnel identifié et que son ROI est calculé et validé

Alors on met en place un plan d'industrialisation

- Code maintenable : code modulaire, séparation back-end/front-end, Docker, robuste aux changements (outils, données d'entrée)
- Pérenniser l'accès aux données : serveurs, automatisation de la collecte et du traitement de données
- Passer de l'algorithme à une application/service (webapp) pour des utilisateurs

3

**ORGANISATION
CHANTIERS DATA**

**DÉVELOPPEMENT AGILE
MÉTHODE SCRUM**

<https://scrumguides.org/>

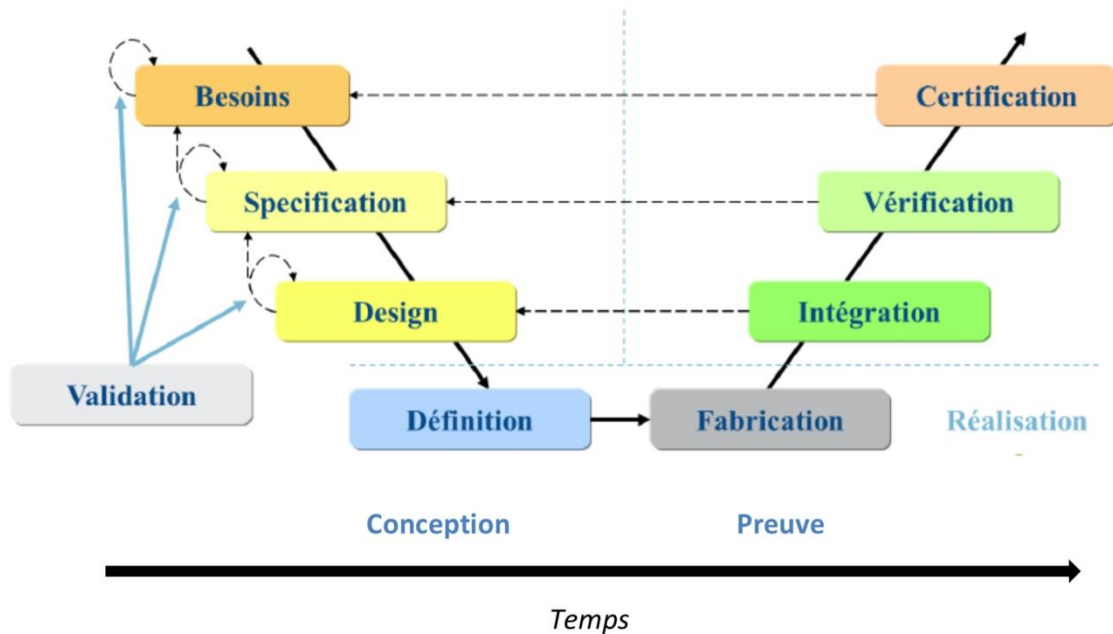
Organisation & règles

Quoi	A Qui	Quand	Resp.	Comment	Où	Commentaires
<u>Flash report</u>	Membres du COFIL	Bi mensuel	Scrum-Owner	Email		
Validation jalon	Membres du COFIL	A la fin du jalon	Scrum-Owner	Réunion	Salle de réunion	Planifier ces réunions en avance pour bloquer les calendriers. N'attendez pas la fin du jalon pour planifier la réunion
Mise en stand-by du chantier	Membres du COFIL	L'arrivée d'un imprévu <u>justifié</u> empêchant le scrum-team d'assurer une disponibilité de 2jrs/sem sur le projet	Scrum-Owner	Email ou Réunion si besoin		Toute indisponibilité du scrum-team doit être officialisée par une communication de mise en stand-by La mise en stand-by du chantier ne doit pas dépasser 3 semaines sinon le chantier sera arrêté définitivement
Arrêt définitif du projet	Equipe projet + membres COFIL	Mise en stand-by dépassant une durée de 1 mois OU Chantier dépassant le délai de 6 mois	Scrum-Master	Réunion avec le data climber et le sponsor	Salle de réunion	Le scrum-master aura la possibilité de representer un nouveau chantier pour valider le projet
Certification du chantier	Comité de Direction + SPONSOR	Session de certification GB	Scrum-Owner	Présentation du chantier 20 min		La certification du projet se fait en même temps que des sessions de certification Green Belt après une présentation de 20 min

Méthode agile (1) - Le cycle en V

Extraits de la thèse d'Estelle Rémondeau (X + Safran) : « De l'adaptation des approches agiles aux systèmes complexes industriels : une mise en œuvre. »

FIGURE I-2 : CYCLE DE DEVELOPPEMENT EN V D'UN SYSTEME COMPLEXE



Méthode agile (2) - Complexité des projets



FIGURE I-4 : LES DIMENSIONS DE LA COMPLEXITE D'UN PROJET
(DE REZENDE & BLACKWELL, 2019)

Méthode agile (3) - Scrum

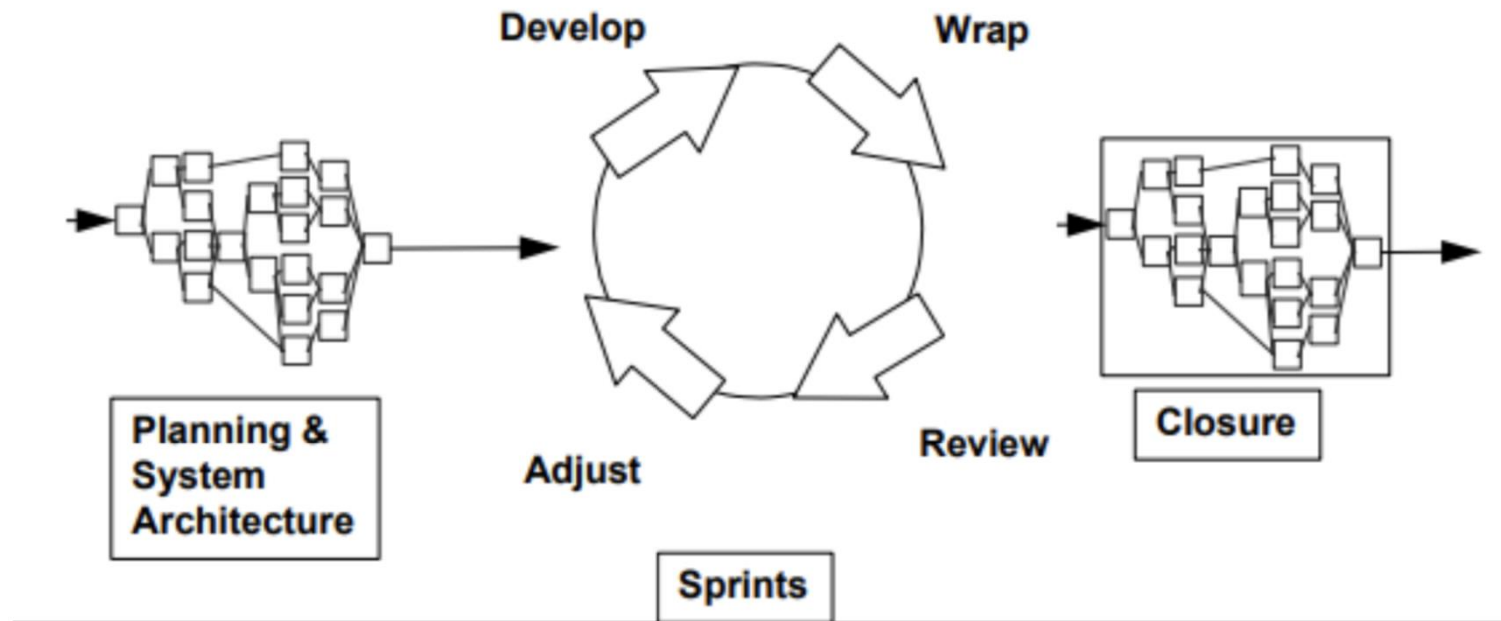


FIGURE II-7 METHODOLOGIE *SCRUM* (SCHWABER, 1997)

Méthode agile (4) - Cycle de vie

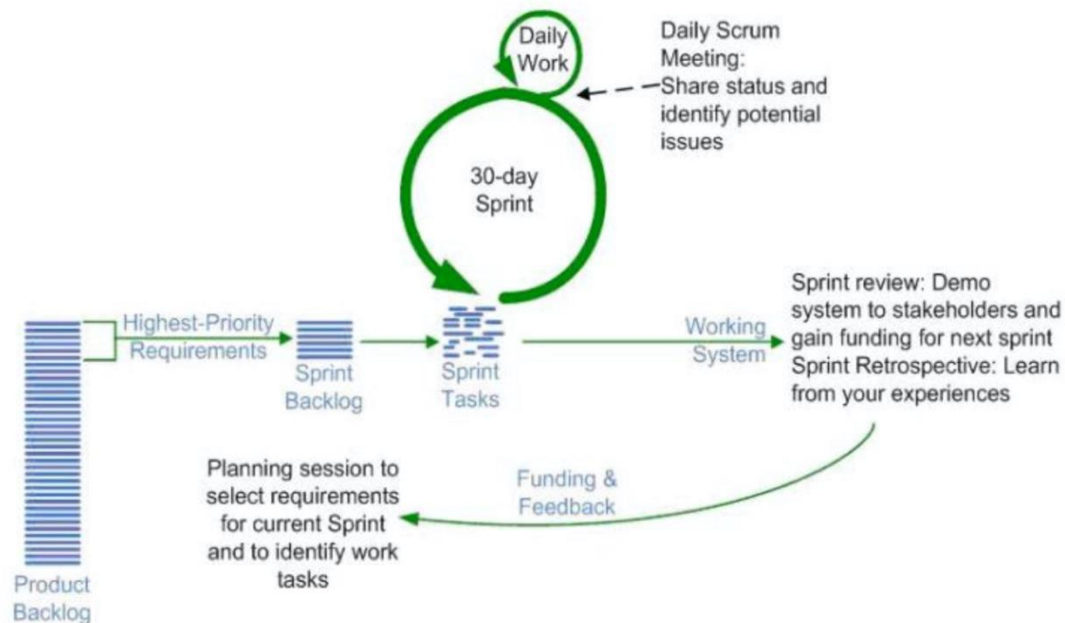


FIGURE II-8 : CYCLE DE VIE *SCRUM* (AMBLER, 2009)

Méthode agile (5) - Décomposition du cycle en V

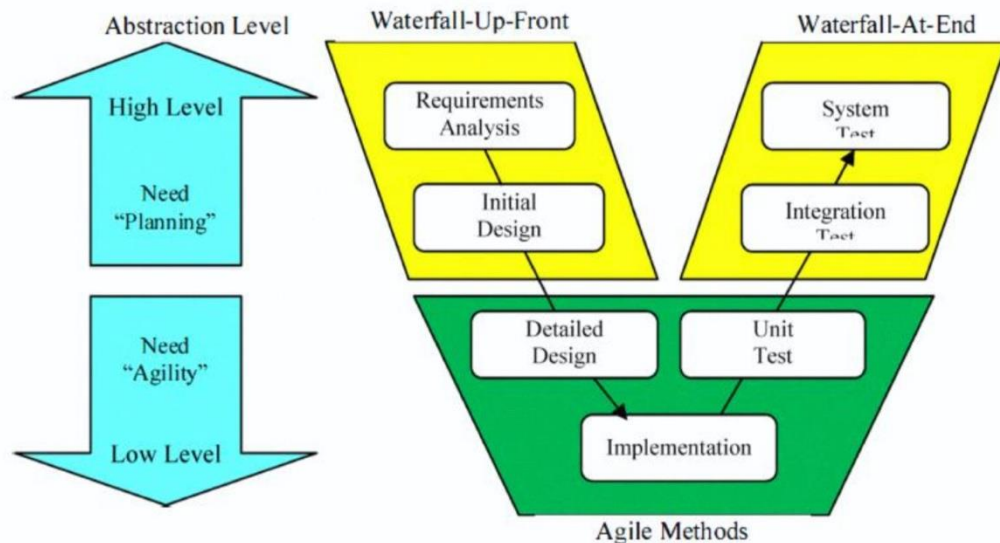


FIGURE II- 9 : UN MODELE DE DEVELOPPEMENT HYBRIDE POUR LE DEVELOPPEMENT LOGICIEL ET LA GESTION DE PROJET (HAYATA & HAN, 2011)

Méthode agile (6) - Cycle de Hype

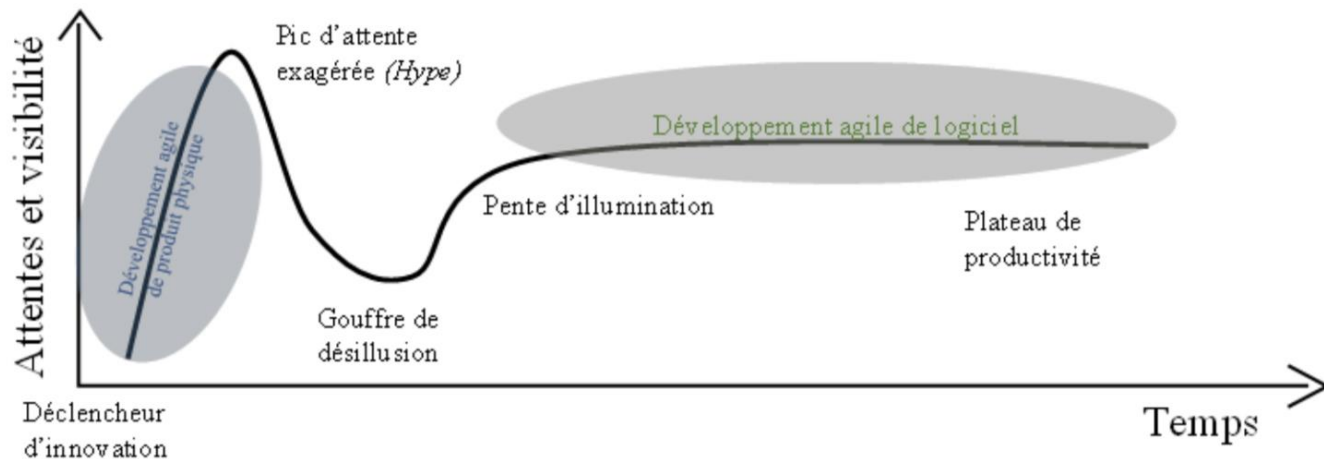


FIGURE III-5 : CYCLE DE HYPE APPLIQUE AU DEVELOPPEMENT AGILE DE LOGICIEL ET AU DEVELOPPEMENT AGILE DE PRODUITS PHYSIQUES (D'APRES SCHMIDT ET AL, 2018)

4

CONCLUSION

Les points à retenir concernant le processus

Cadrage

- Pensez à bien cadrer le sujet de votre chantier (quel est le problème, quel est le plan pour le résoudre...)
- Ajuster votre plan de travail en fonction du délai du chantier (6 mois max avec 2 jrs/sem)

Nettoyage des données

- Avant de commencer à nettoyer les données brutes
- Sélectionner les variables avec l'aide des métiers
- Sélectionner un échantillon avec un nombre d'observations suffisant pour le projet

Analyse descriptive

- Elle permet de découvrir des tendances et de donner des indications pour la modélisation, ne pas se précipiter à faire des modèles.
- Cette étape doit être guidée par un objectif opérationnel

Rentabilité et efficacité

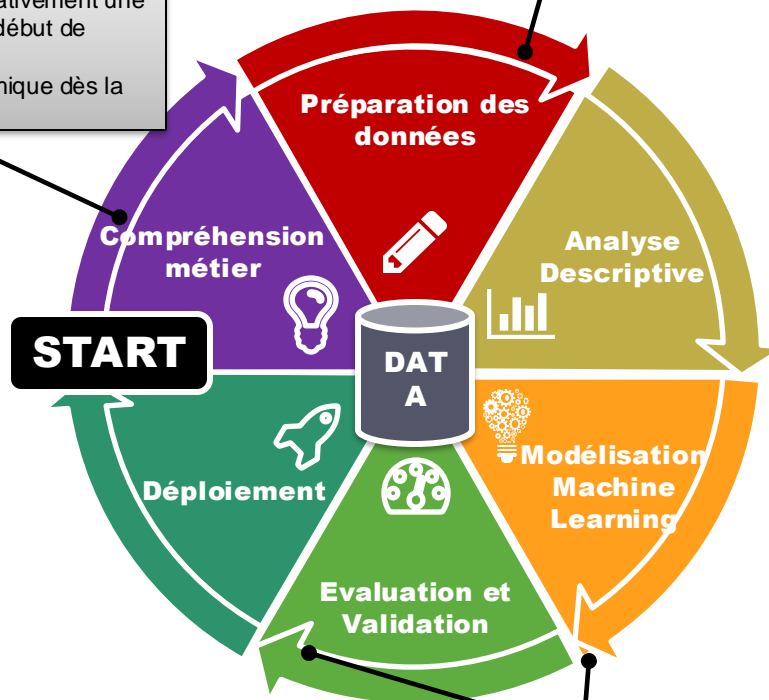
- Réfléchir au ROI
- Sortir rapidement en cas d'échec

Sortir si on ne parvient pas à :

- Identifier au moins qualitativement une utilisation du modèle au début de projet
- Quantifier le gain économique dès la découverte d'un modèle

Sortir si on ne parvient pas à :

- Collecter un minimum les données pertinentes
- Parvenir à des données propres



Sortir si on ne parvient pas à :

- Trouver un modèle performant



QUESTIONS ?