DATA MINING

Les méthodologies de travail

4 ERP-BI-4TWIN-4SAE-4INFINI
2023-2024

Equipe Data Mining

KDD / ECD

KNOWLEDGE DATA DISCOVERY
EXTRACTION DE CONNAISSANCES À PARTIR DE DONNÉES

DÉFINITION

PHASES PRINCIPALES

APPLICATION

esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

KDD: DÉFINITION

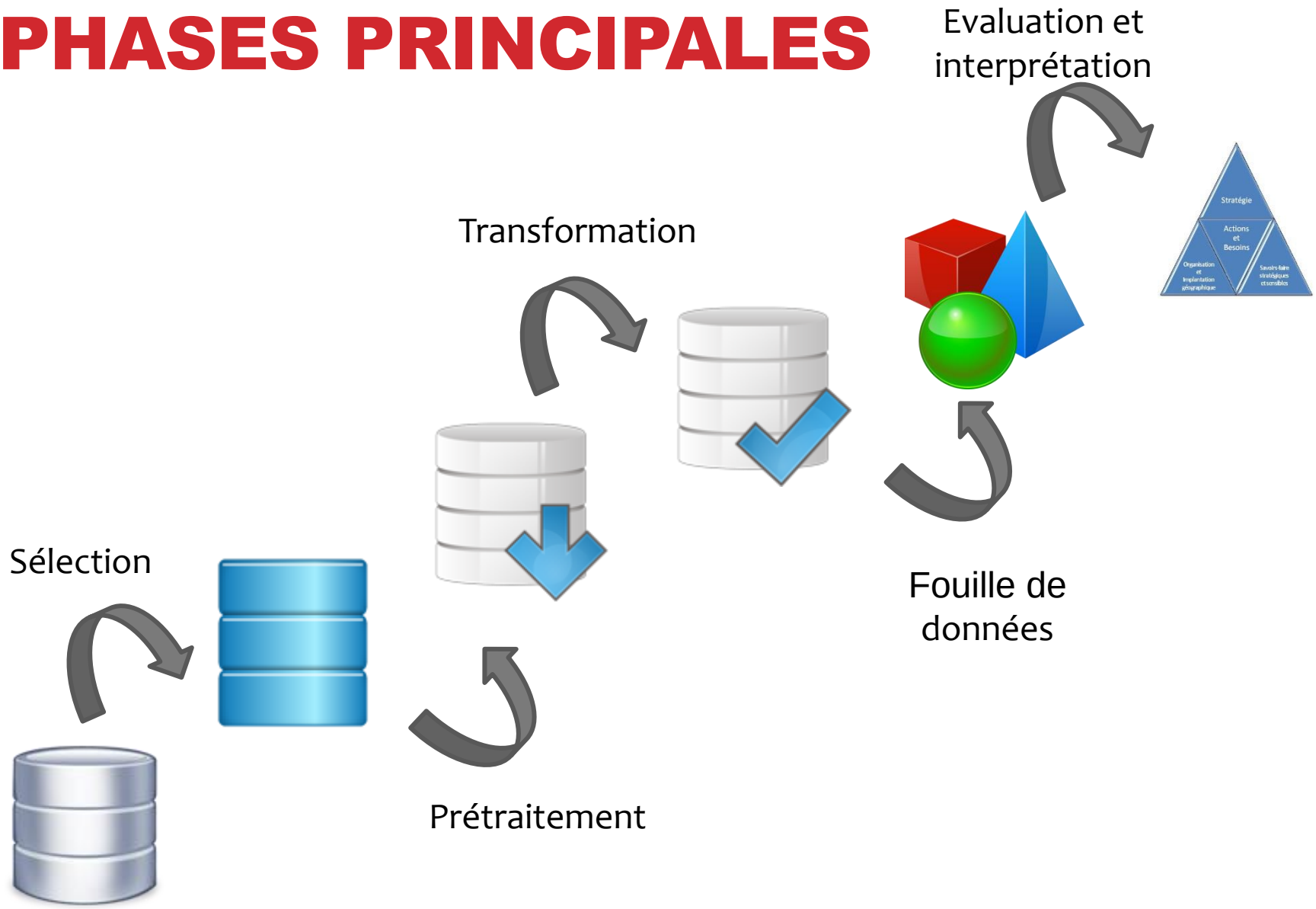
Knowledge Discovery in Databases

- proposé par Ossama Fayyad en 1996
- un processus pour la fouille de données qui a bien répondu aux besoins d'entreprises, et qui est devenu rapidement très populaire.
- KDD a comme but l'extraction des connaissances,
- des motifs valides, utiles et exploitables à partir des grandes quantités de données
- par des méthodes automatiques ou semi-automatiques.

KDD: DÉFINITION

- Le processus de KDD est **itératif** et **interactif**.
- Le processus est itératif : il peut être nécessaire de refaire les pas précédents.
- Le problème de ce processus, comme pour les autres présentés dans la section suivante, est le manque de guidage de l'utilisateur, qui ne choisit pas à chaque étape la meilleure solution adaptée pour ses données.

KDD: PHASES PRINCIPALES



KDD:
PHASES PRINCIPALES

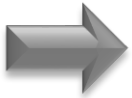
1. Développer et comprendre le domaine de l'application

C'est le pas initial de ce processus. Il prépare la scène pour comprendre et développer les buts de l'application.

KDD:
PHASES PRINCIPALES

2. Sélection des données

La sélection et la création d'un ensemble de données sur lequel va être appliqué le processus d'exploration.



Données ciblées

KDD:
PHASES PRINCIPALES

3. Le prétraitement et le nettoyage des données

Cette étape inclut des opérations comme l'enlèvement du bruit et des valeurs aberrantes -si nécessaire, des décisions sur les stratégies qui vont être utilisées pour traiter les valeurs manquants...

Données prétraitées

KDD:
PHASES PRINCIPALES

4. La transformation des données

Cette étape est très importante pour la réussite du projet et doit être adaptée en fonction de chaque base de données et des objectifs du projet.

Dans cette étape nous cherchons les méthodes correctes pour représenter les données. Ces méthodes incluent la réduction des dimensions et la transformation des attributs.

➡ **Données transformées**

- Une fois que toutes ces étapes seront terminées, les étapes suivantes seront liées à la partie de Data mining, avec une orientation sur l'aspect algorithmique.

KDD:
PHASES PRINCIPALES

5. Choisir la meilleure tâche pour Datamining

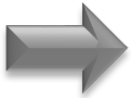
Nous devons choisir quel type de Datamining sera utilisé, en décidant le but du modèle.

- Par exemple : classification, régression, regroupement...

KDD:
PHASES PRINCIPALES

6. Choisir l'algorithme de Datamining

Dans cette étape nous devons choisir la méthode spécifique pour faire la recherche des motifs, en décidant quels modèles et paramétrés sont appropriés.



Modèles

KDD:
PHASES PRINCIPALES

7. Implémenter l'algorithme de Datamining

Dans cette étape nous implémentons les algorithmes de Datamining choisis dans l'étape antérieure.

Peut être il sera nécessaire d'appliquer l'algorithme plusieurs fois pour avoir le résultat attendu.

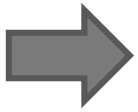
KDD:
PHASES PRINCIPALES

8. Evaluation

Evaluation et interprétation des motifs découverts.

Cette étape donne la possibilité de:

- Retourner à une des étapes précédentes
- Avoir une représentation visuelle des motifs, enlever les motifs redondants ou non-représentatifs et les transformer dans des termes compréhensibles pour l'utilisateur.

 **Connaissances**

KDD:
PHASES PRINCIPALES

9. Utiliser les connaissances découvertes

Incorporation de ces connaissances dans des autres systèmes pour d'autres actions.

Nous devons aussi mesurer l'effet de ces connaissances sur le système, vérifier et résoudre les conflits possibles avec les connaissances antérieures.

KDD: APPLICATION

Le KDD est devenu lui-même un modèle pour les nouveaux modèles.

Le modèle a été utilisé dans plusieurs domaines différentes : ingénierie, médicine, e-business, production, développement du logiciel, etc.

CRISP-DM

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

NCR

SPSS

Mercedes-Benz

DÉFINITION

PHASES PRINCIPALES

EXEMPLE

esprit

Ecole Supérieure Privée
d'Ingénierie et de Technologies

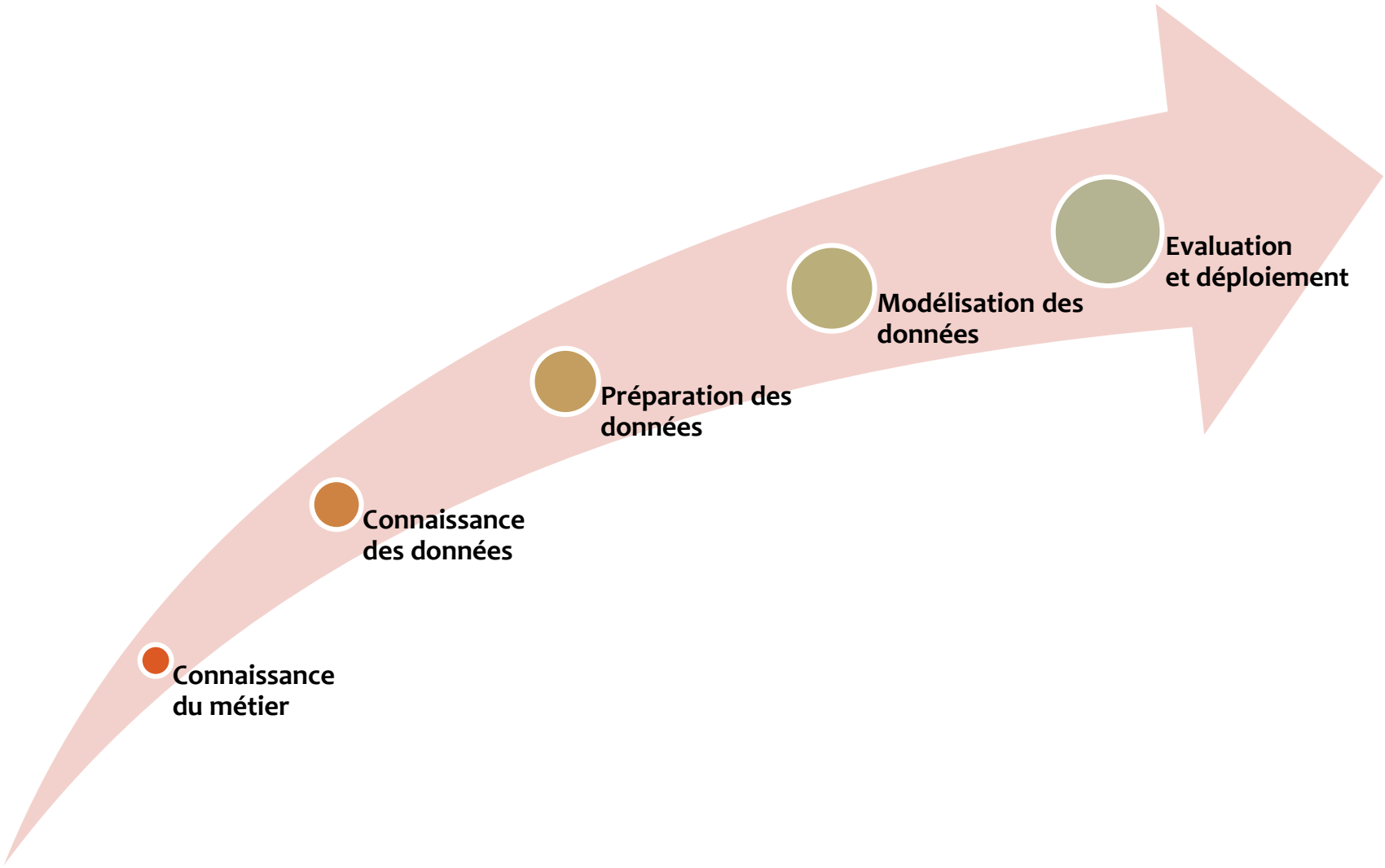
CRISP-DM : DÉFINITION

Cross-Industry Standard Process for Data Mining

Une méthode mise à l'épreuve sur le terrain permettant d'orienter les travaux de Data mining

Processus de data mining qui décrit une approche communément utilisée par les experts pour résoudre les problèmes qui se posent à eux.

PHASES PRINCIPALES



CRISP-DM : DÉFINITION

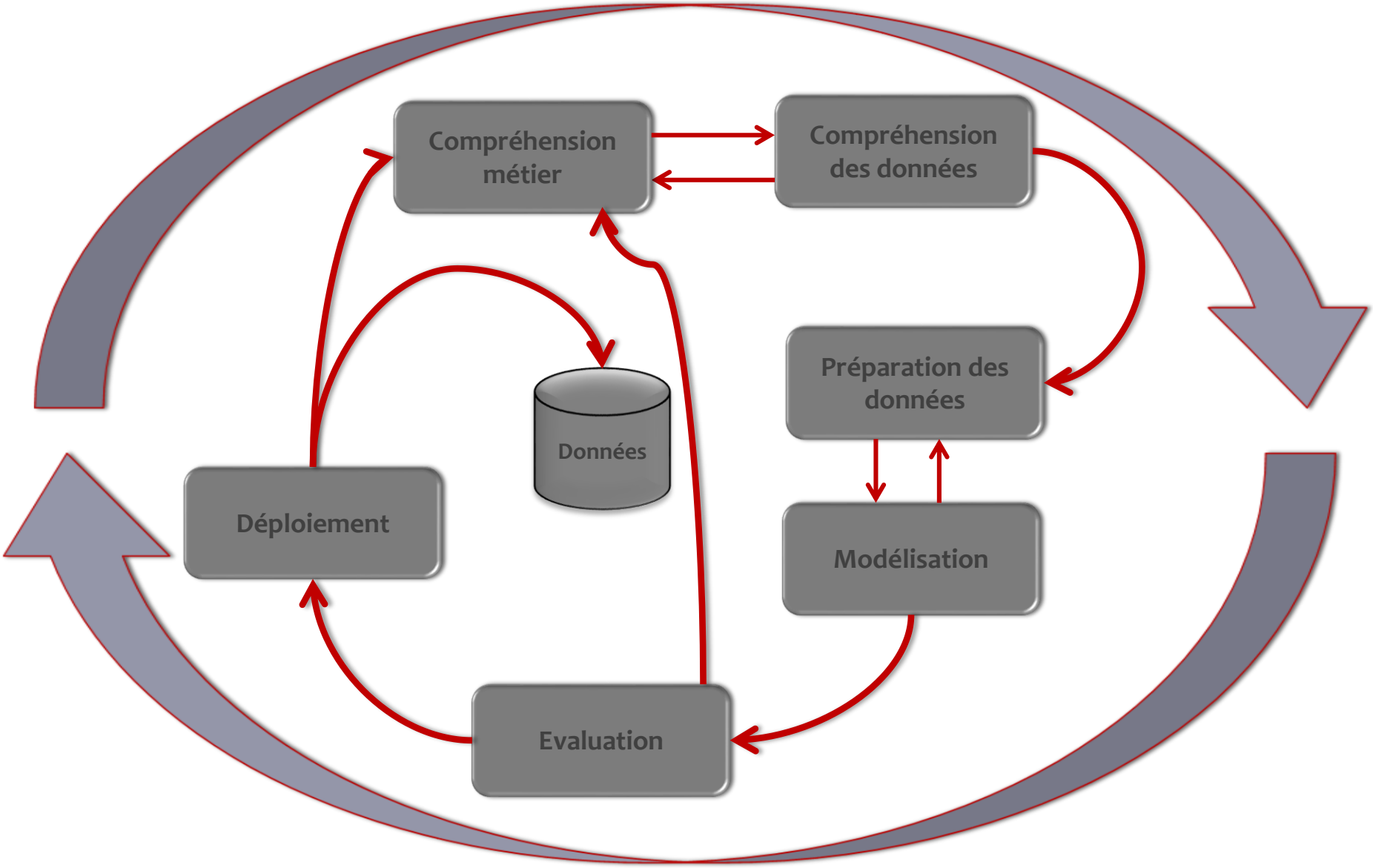
Méthodologie

- ✓ comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.

Modèle de processus

- ✓ offre un aperçu du cycle de vie du Data mining.

CRISP-DM : PHASES PRINCIPALES



1. COMPRÉHENSION MÉTIER

Déterminer les objectifs d'affaires

Résoudre un problème spécifique

Evaluer la situation actuelle

Convertir en un problème de data mining

- ✓Quels types de clients sont intéressés par chacun de nos produits?
- ✓Quels sont les profils typiques de nos clients?

Élaborer un plan de projet

2. COMPRÉHENSION DES DONNÉES

Collecte de données initiale

Description des données

Exploration des données

Vérification de la qualité des données

Sélection des données

Les données connexes peuvent provenir de nombreuses sources :

- ✓Interne (ERP, CRM, Data Warehouse…)
- ✓Externe (données commerciales, données du gouvernement…)
- ✓Créées (recherche)

LES ENJEUX DE LA SÉLECTION DES DONNÉES

Mettre en place une description concise et claire du problème

- ✓Identifier les comportements de dépenses des femmes qui achètent des vêtements saisonniers
- ✓Identifier les modèles de la faillite de détenteurs de cartes de crédit

Identifier les données pertinentes pour la description du problème

- ✓Données démographiques, données financières…

Les variables sélectionnées pour les données pertinentes doivent être indépendantes les unes des autres.

3. PRÉPARATION DES DONNÉES

Nettoyer les données sélectionnées pour une meilleure qualité

- ✓Remplissez les valeurs manquantes
- ✓Identifier ou supprimer les valeurs aberrantes
- ✓Résoudre la redondance causée par l'intégration des données
- ✓Les données incohérentes correctes

Transformer les données

- ✓Convertir des mesures différentes de données dans un échelle numérique unifié en utilisant des formulations mathématiques simples

LES DONNÉES DANS LE MONDE RÉEL!

Incomplètes: manque de valeurs d'attributs, manque de certains attributs d'intérêt ou ne contenant que des agrégats des attributs d'intérêt ou contenant uniquement des données agrégées

- ✓ l'occupation = ""

Bruyantes: contenant des erreurs ou des valeurs aberrantes

- ✓ Salaire = "- 1000"

Incompatibles: contenant des écarts dans les codes ou les noms

- ✓ Age = "42" et anniversaire = "03/07/1993"
- ✓ note =« 1,2,3 » ensuite «A, B, C »

PRINCIPALES CAUSES

Les données incomplètes peuvent provenir de:

- ✓La valeur de données lors de la collecte «Sans objet»
- ✓Des considérations différentes entre le moment où les données ont été collectées et lorsqu'elles sont analysées.
- ✓Problèmes humains / matériels / logiciels

Les données bruyantes (valeurs incorrectes) peuvent provenir de:

- ✓Les instruments de collecte de données sont erronés
- ✓L'erreur humaine ou informatique à la saisie de données
- ✓Les erreurs de transmission de données

Les données incohérentes peuvent provenir de:

- ✓Les différentes sources de données
- ✓La violation de la dépendance fonctionnelle (par exemple, de modifier certaines données liées)

TRANSFORMATION DES DONNÉES

Transformez le numérique à des échelles numériques

- ✓Les échelles salariales de « 100 TND » à « 1000 TND » à un certain nombre de [0.0, 1.0]
- ✓Le système métrique (par exemple, le mètre, kilomètre) au système anglais (par exemple, des pieds et miles)

Recoder les données catégoriques à des échelles numériques

- ✓"1" = "oui" et "0" = "No"

4. MODÉLISATION

Traitement des données

- ✓Ensemble d'apprentissage
- ✓Ensemble de test…

Les techniques de data mining

- ✓Association
- ✓Classification
- ✓Clustering
- ✓Prédictions
- ✓Les motifs séquentiels

5. EVALUATION

Est-ce que le modèle répond aux objectifs métier?

Des objectifs métier importants non résolus?

Est-ce que le modèle est logique?

Est-ce que le modèle est actionnable?

Il devrait être possible de prendre des décisions après cette étape.

Tous les objectifs importants doivent être atteints.

6. DÉPLOIEMENT / IMPLÉMENTATION

En cours de suivi et d'entretien

- ✓Évaluer la performance par rapport aux critères de réussite
- ✓La réaction du marché et les changements des concurrents

CRISP-DM:
ETUDE DES FACTURES DE TÉLÉPHONE

Problème : Les factures de téléphone non payées.

- Le data mining utilisé pour développer des modèles pour prédire le non paiement des factures au plus tôt possible.

CRISP-DM: Exemple
Etude des factures de téléphone

Séquence de période de facturation:

- ✓Utilisez 2 mois, recevoir la facture, le paiement du mois de facturation, débrancher si la facture n'est pas réglée pendant une période déterminée

1. COMPRÉHENSION MÉTIER

Prédire quels clients seraient insolvables

- ✓ À temps pour l'entreprise pour prendre des mesures préventives (et d'éviter de perdre de bons clients)

hypothèse:

- ✓ Clients insolvables vont changer les habitudes d'appel et l'usage du téléphone pendant une période critique avant et immédiatement après la fin de la période de facturation.

2. COMPRÉHENSION DES DONNÉES

Les informations statiques des clients sont disponibles dans des fichiers

- ✓ Factures, paiements, utilisation…

Un entrepôt de données est utilisé pour recueillir et organiser les données

- ✓Un codage pour protéger la vie privée des clients

CRÉATION DE L'ENSEMBLE DES DONNÉES CIBLES

Les fichiers des clients:

- ✓Informations sur les clients
- ✓Déconnexion
- ✓Reconnexions

Données dépendantes du temps

- ✓Factures
- ✓Paiements
- ✓Utilisation

100, 000 clients sur une période de 17 mois

L'échantillonnage pour assurer à tous les groupes une représentation appropriée

3. PRÉPARATION DES DONNÉES

Filtrer les données incomplètes

Les appels en promotion supprimés

- ✓Le volume des données réduit d'environ 50%

Faible nombre des cas de fraude

Vérification croisée avec les déconnexions du téléphone

Les données retardées sont nécessairement synchronisées

5. MODÉLISATION

Analyse discriminante

- ✓Le modèle linéaire

Les arbres de décision

- ✓Classificateur à base de règles

Réseaux de Neurones

- ✓Le modèle non linéaire

5. EVALUATION

Premier objectif est de maximiser la précision de la prédiction des clients insolvables

- ✓Arbre de décision un classificateur meilleur

Deuxième objectif est de minimiser le taux d'erreur pour les clients de solvants

- ✓Le modèle Réseau de Neurones proche de l'arbre de décision

Utilisé tous les 3 sur la base de cas par cas.

6. IMPLÉMENTATION

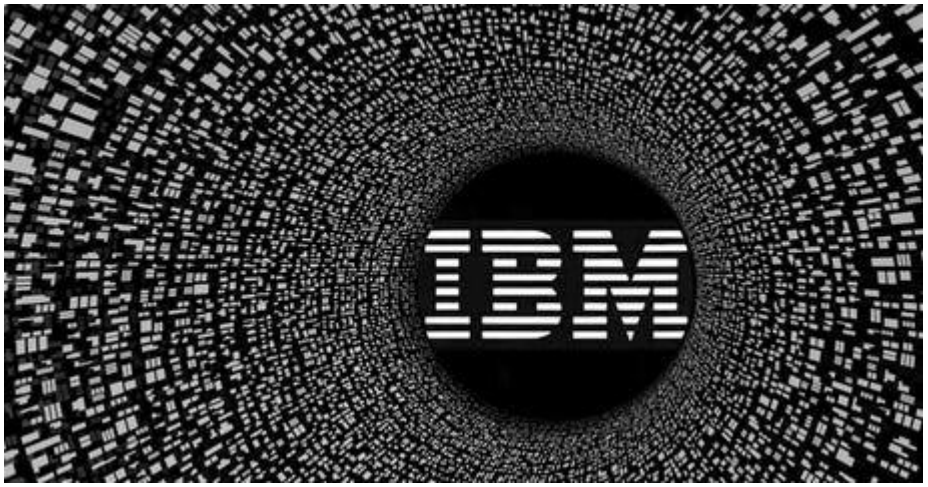
Chaque client a été examiné avec les 3 algorithmes

- ✓Si tous les 3 sont convenables, utiliser une classification
- ✓En cas de désaccord, catégorisé comme non classé

Correcte sur les données d'essai avec 0.898

- ✓Seulement 1 client solvant aurait été débranché

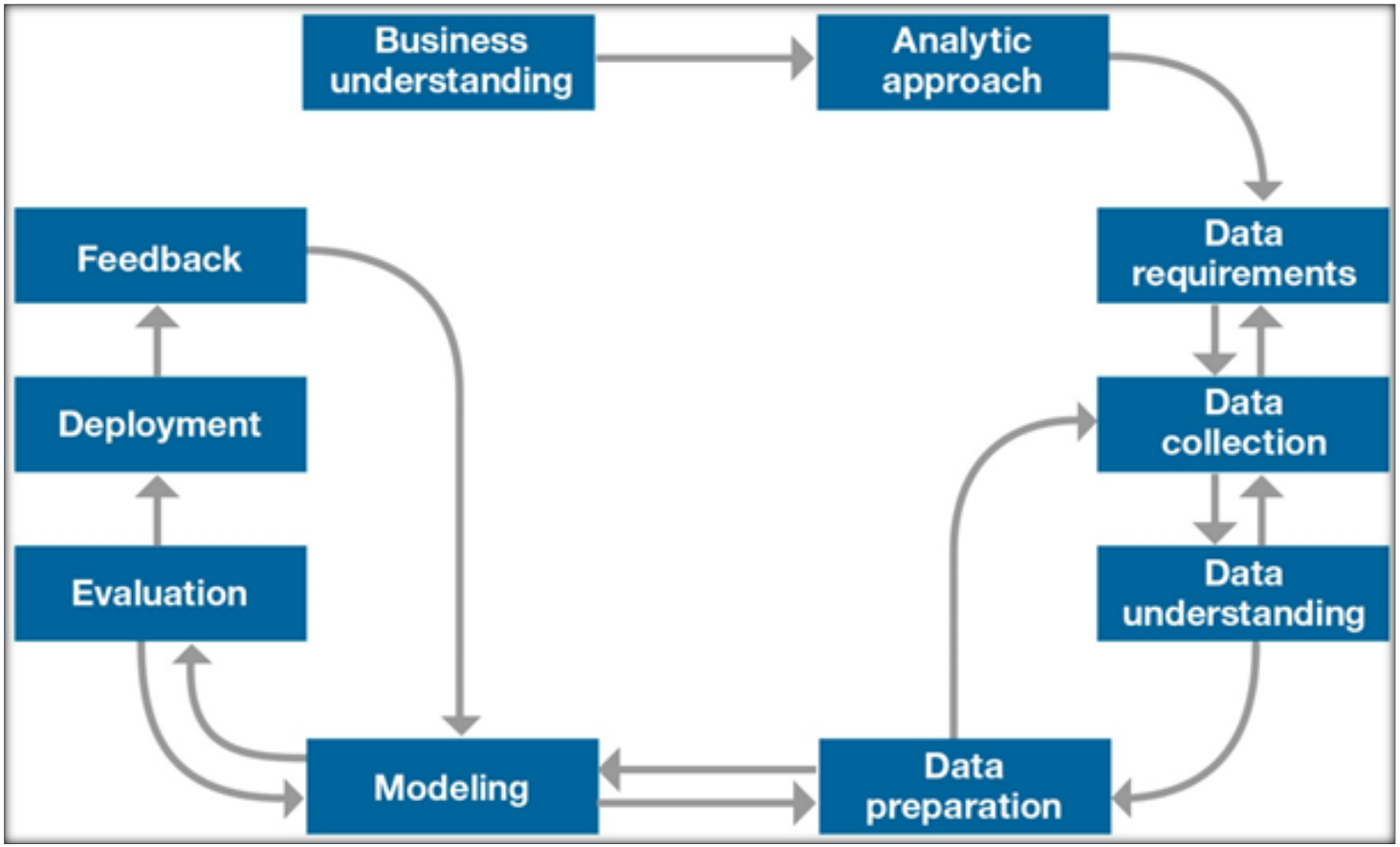
IBM MASTER PLAN



PRINCIPALES PHASES

esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

IBM MASTER PLAN : PHASES PRINCIPALES



PHASE 1: COMPRÉHENSION DU MÉTIER

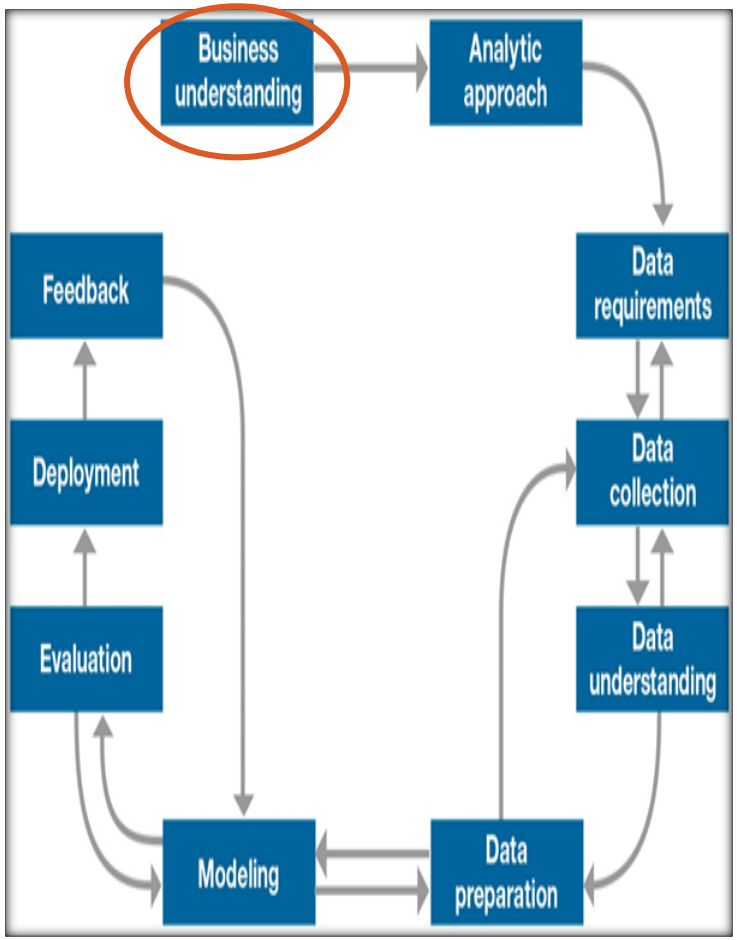
Chaque projet, quelle que soit sa taille, commence par la compréhension des activités de l'entreprise.

- Le rôle essentiel de cette étape est:

- définir le problème

- identifier les objectifs métiers et les exigences de la solution d'un point de vue commercial.

- Cette première étape est la plus difficile.

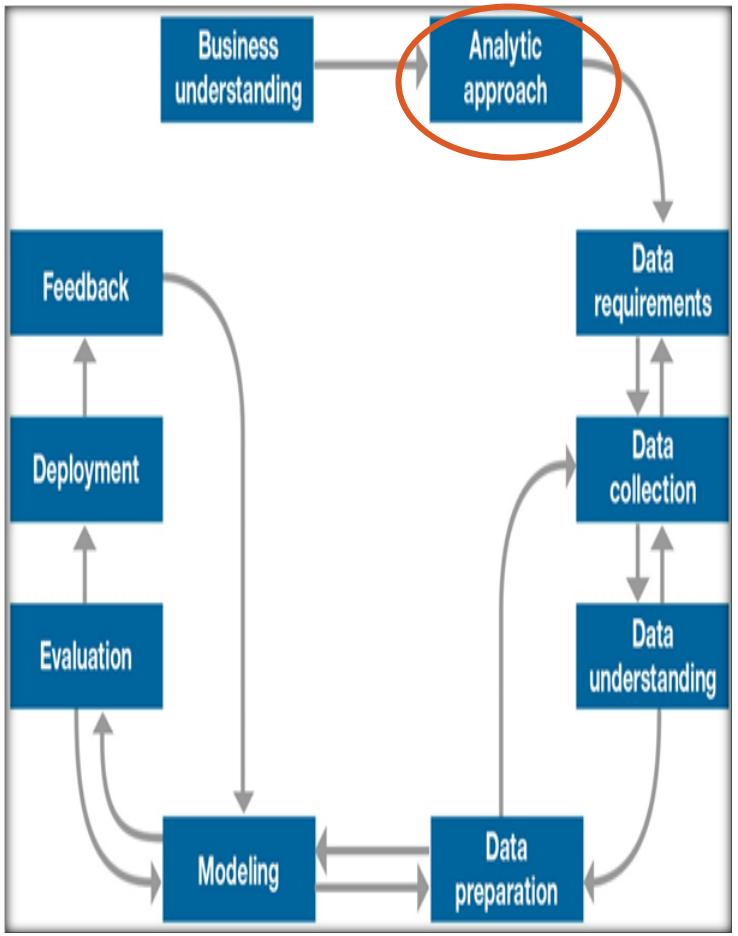


PHASE 2: APPROCHE ANALYTIQUE

- Après avoir clairement énoncé le problème commercial, le 'spécialiste des données' peut définir l'approche analytique permettant de le résoudre.

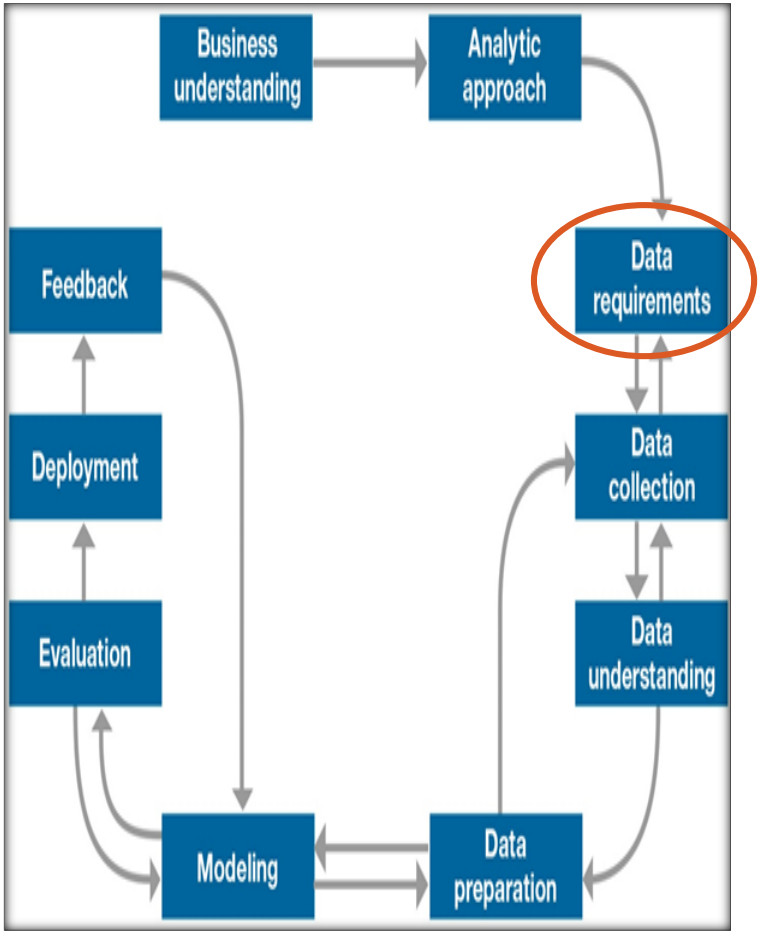
→ Définir les objectifs data Science;

→ identifier les techniques de Data Mining et de Machine Learning permettant d'atteindre le résultat souhaité.



PHASE 3: EXIGENCES NIVEAU DONNÉES

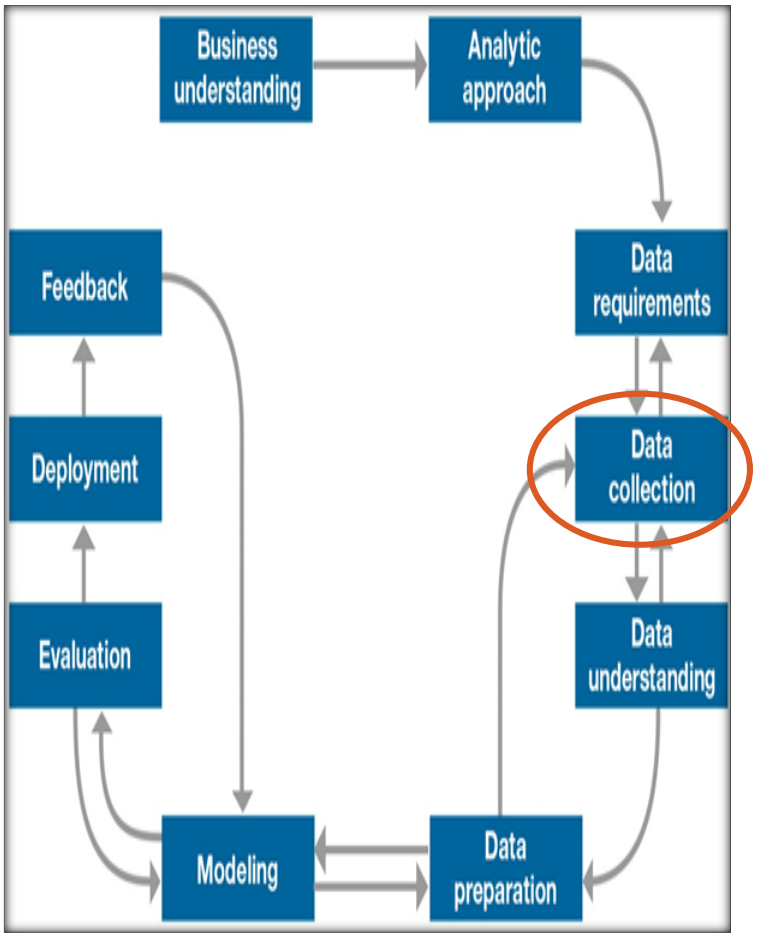
- Le choix de l'approche analytique détermine les exigences en matière de données;
- Les méthodes analytiques à utiliser nécessitent un contenu, des formats et des représentations de données particuliers, guidés par les connaissances du domaine.



PHASE 4: COLLECTE DES DONNÉES

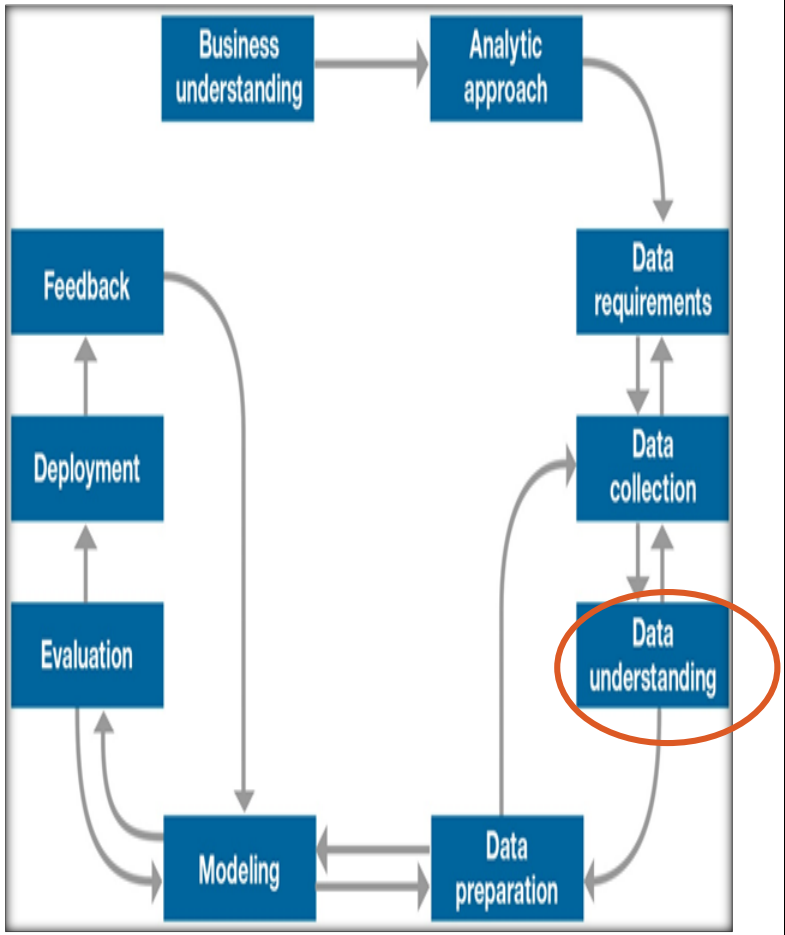
- Cette étape consiste à identifier et rassembler des ressources de données - structurées, non structurées et semi-structurées – qui sont pertinentes pour le problème.

- Lorsqu'il rencontre des lacunes dans la collecte de données, le 'spécialiste des données' peut avoir besoin de réviser les exigences en matière de données et de collecter davantage de données.



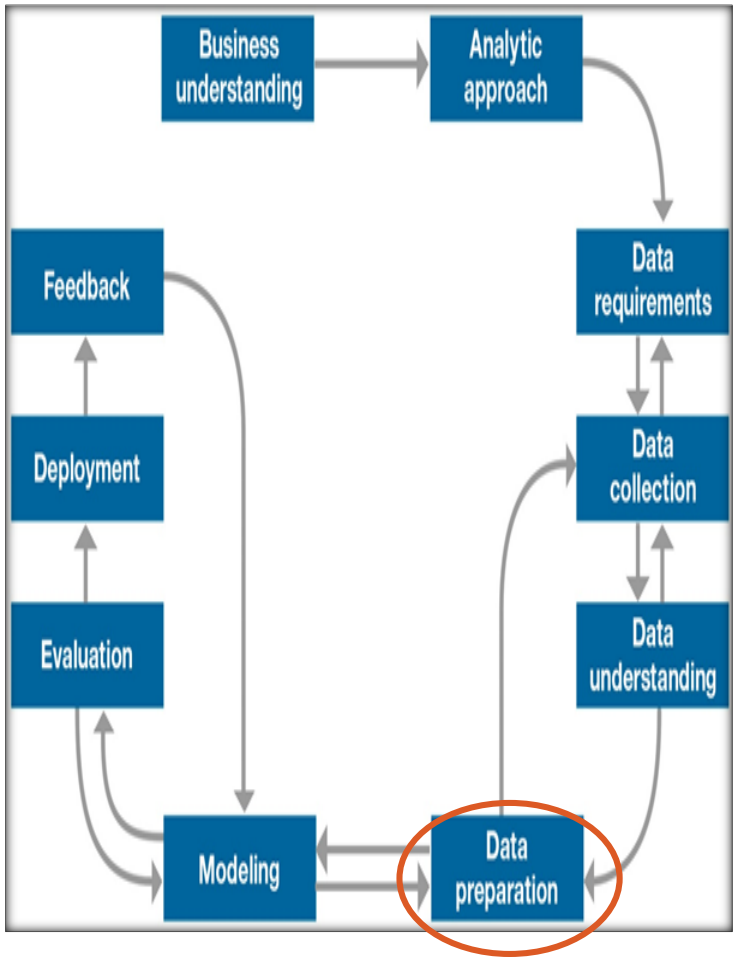
PHASE 5: COMPREHENSION DES DONNÉES

- Des statistiques descriptives et des techniques de visualisation peuvent aider un scientifique des données à comprendre le contenu des données, à évaluer leur qualité et à découvrir les informations initiales relatives à ces données.
- La collecte de données, pourrait être nécessaire pour combler les lacunes de compréhension.



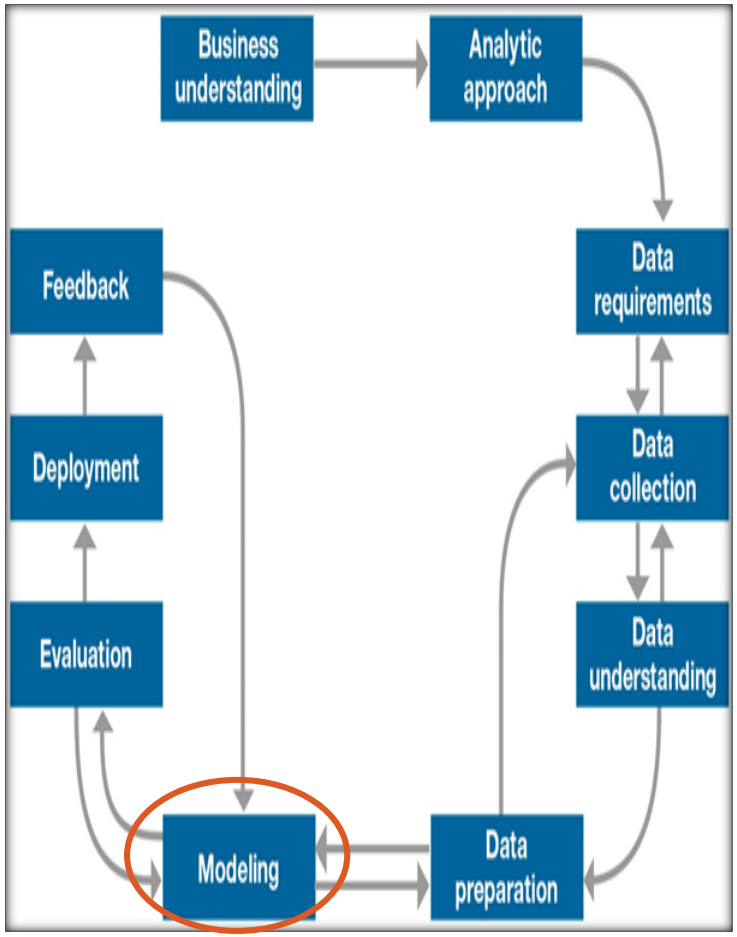
PHASE 6: PRÉPARATION DES DONNÉES

- L'étape de préparation des données comprend toutes les activités utilisées pour construire l'ensemble de données utilisées à l'étape de modélisation.
- Elle inclue le nettoyage des données, la combinaison de données provenant de sources multiples et la transformation de données en variables plus utiles.
- L'ingénierie des caractéristiques ('Feature engineering' et l'analyse de texte peuvent être utilisées pour dériver de nouvelles variables structurées, enrichissant ainsi l'ensemble des prédicteurs et améliorant la précision du modèle.
- C'est l'étape la plus longue (elle représente 70% de la durée totale du projet)



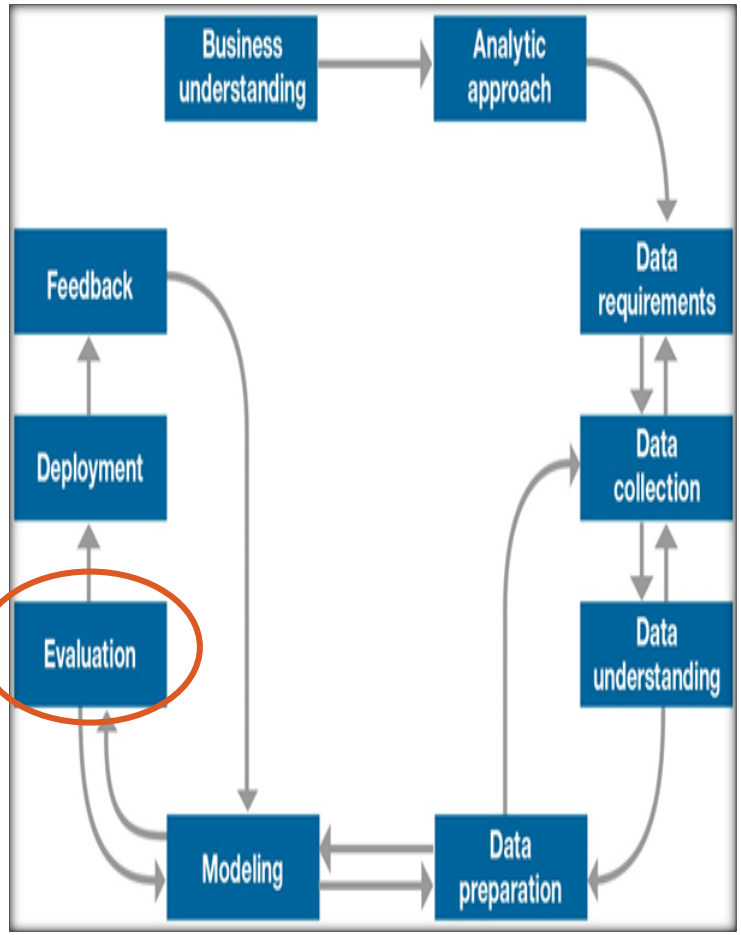
PAHSE 7: MODELISATION DES DONNÉES

- À partir de la première version de l'ensemble de données préparé, les scientifiques utilisent un ensemble d'apprentissage- des données historiques dans lesquelles le résultat recherché est connu -pour développer des modèles prédictifs ou descriptifs à l'aide de l'approche analytique déjà décrite.
- Le processus de modélisation est très itératif



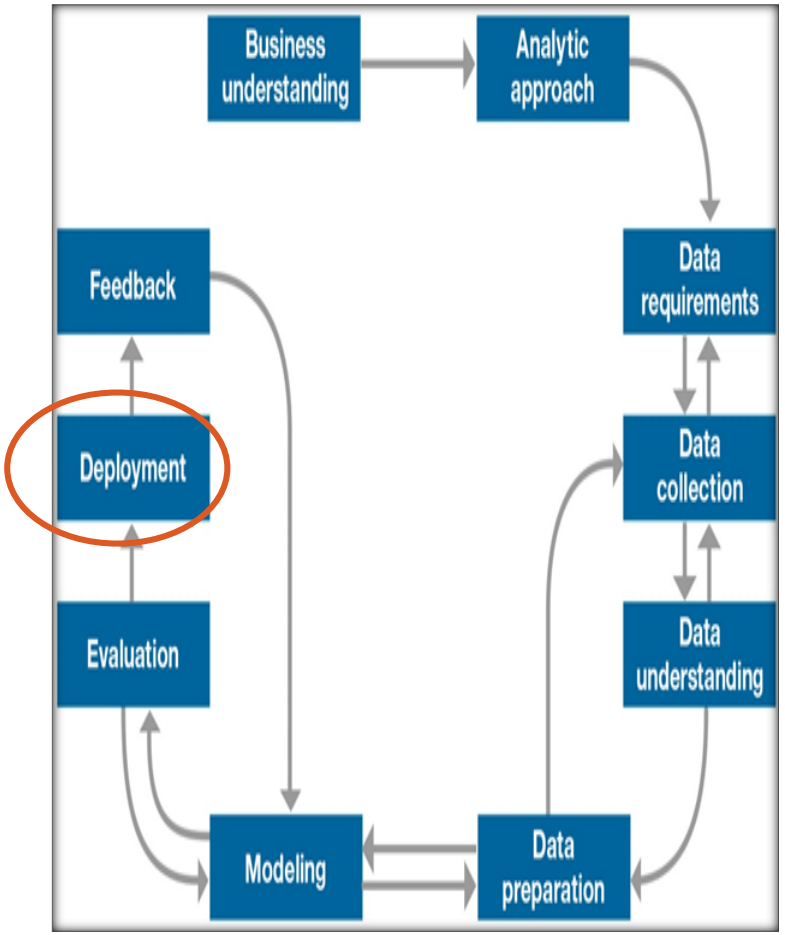
PHASE 8: EVALUATION DES DONNÉES

- L'informaticien évalue la qualité du modèle et vérifie si celui-ci résout le problème de manière complète et appropriée.
- Calculer diverses mesures de diagnostic, ainsi que d'autres résultats, tels que des tableaux et des graphiques, à l'aide d'un ensemble de tests pour un modèle prédictif.



PHASE 9: DÉPLOIEMENT

- Une fois qu'un modèle satisfaisant a été développé et approuvé par les sponsors commerciaux, il est déployé dans l'environnement de production ou dans un environnement de test comparable.
- Le déploiement d'un modèle dans un processus métier opérationnel implique généralement plusieurs groupes, compétences et technologies.



PHASE 10: RETOURS DES UTILISATEURS

- En collectant les résultats du modèle mis en œuvre, l'organisation reçoit des informations en retour sur les performances du modèle et observe son impact sur son environnement de déploiement.
- L'analyse de ces commentaires permet au 'spécialiste des

données' d'affiner le modèle, en augmentant sa précision et donc son utilité.