Les enjeux méthodologiques en Data Science

Mohamed Heny SELMI

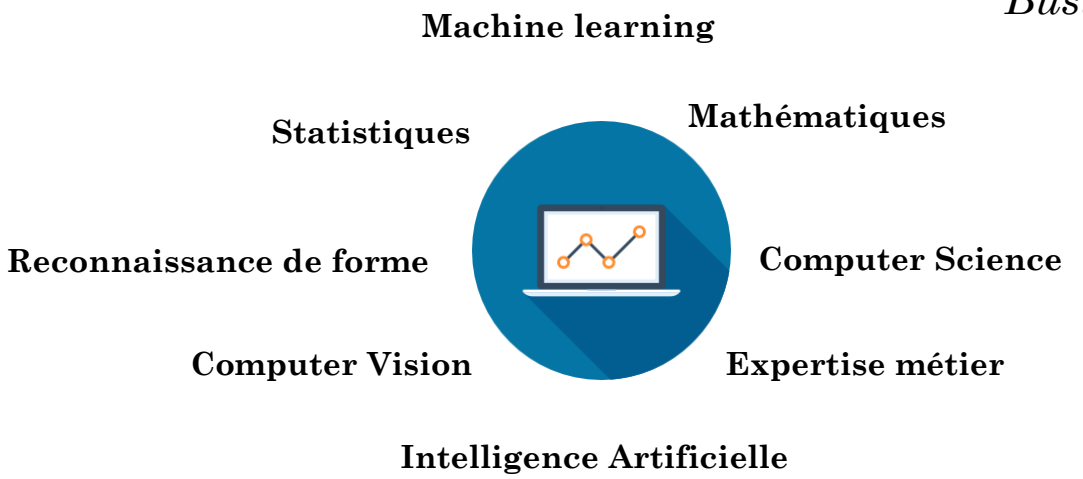
Définitions de la Data Science

Discipline cherchant à extraire de l'intelligence des données, dans le but de la rendre actionnable par les métiers, en s'appuyant principalement sur la Statistique et le Machine Learning, et en utilisant des techniques qui ne sont pas accessible ni par la BI traditionnelle, la DataViz ou l'exploration de données.

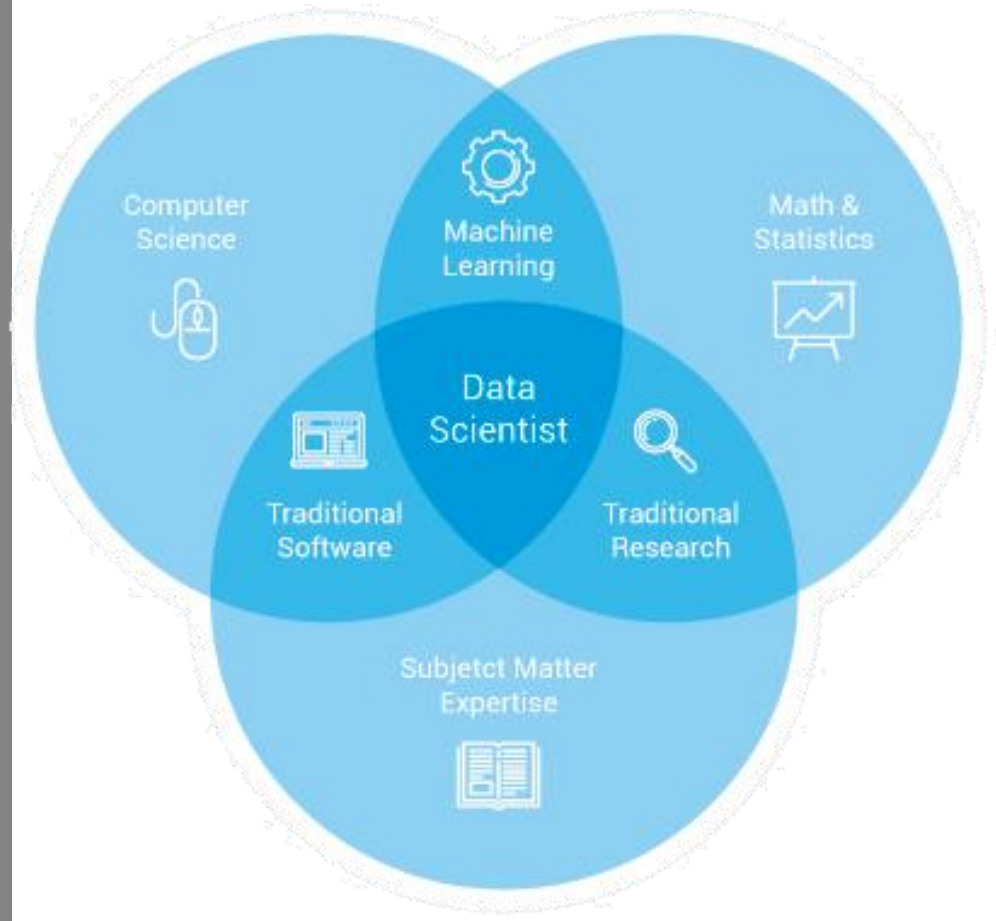
Didier Gaultier
Directeur Data Science
Business & Decision

"Démarche empirique qui se base sur des données pour apporter une réponse à des problèmes »

Data science : fondamentaux et études de cas, E. Biernat, M. Lutz, Eyrolles, 2015.



Aspect pluridisciplinaire



Des origines culturelles très variées :

- Statistiques
- Mathématiques
- Machine Learning
- IT et Programmation
- Techniques de Reporting et Data Viz
- Big Data
- **Métier**

Une approche agile et itérative

- Versus une approche projet classique

Une évolution extrêmement rapide des technologies

L'arrivée massive de l'open source

- Culture open source versus éditeurs

L'arrivée en force du Big Data

Statistiques

Ere du produit

Honorer la «promesse » produit

- Piloter la performance
- Maîtrise des processus
- Optimiser le time-to-market

Data Mining

Ere du client

Conquérir les clients

- Déployer la CRM
- Compléter le marketing produit avec le marketing client

Data Science

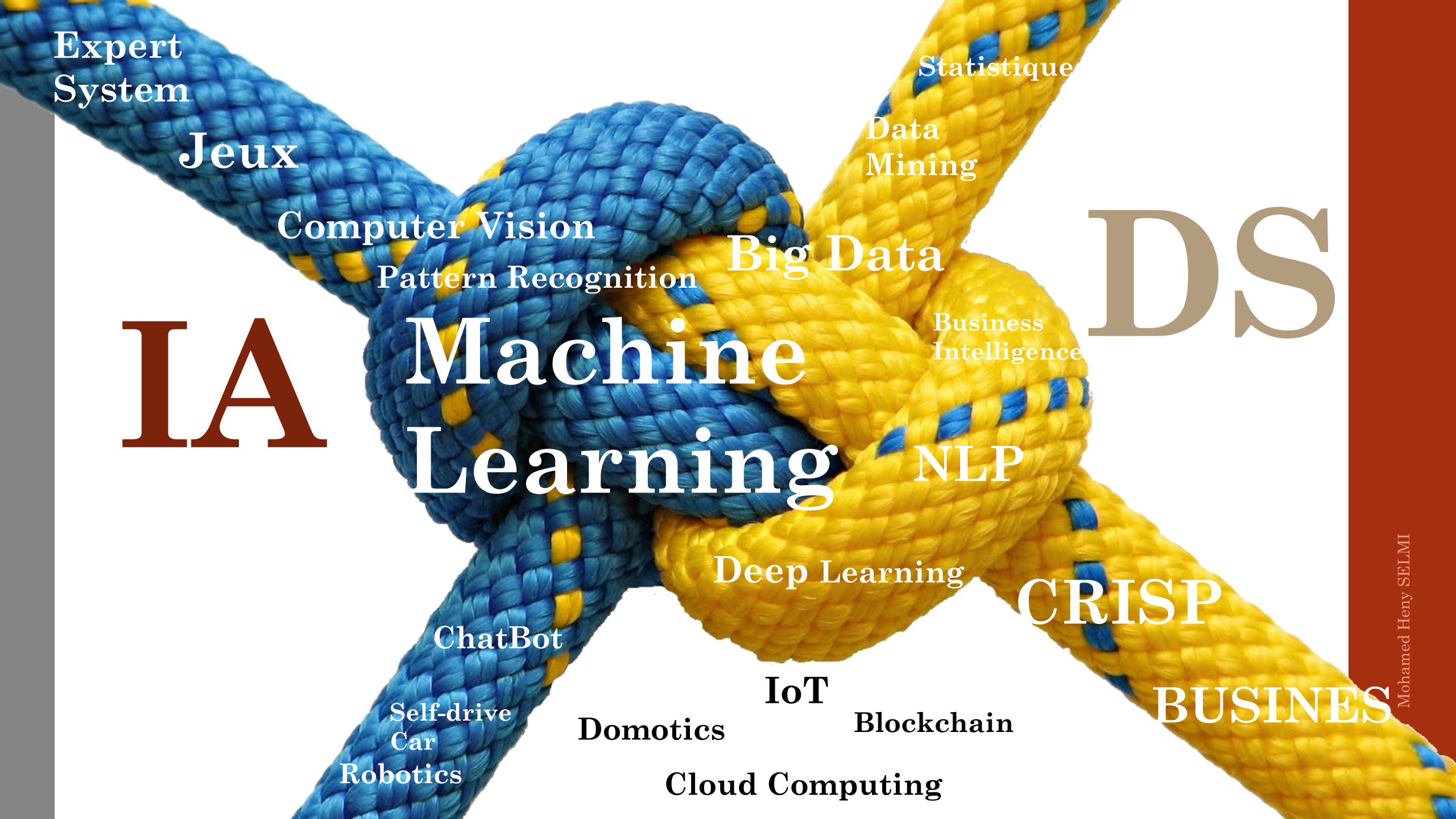
Ere de la donnée

La donnée au cœur des processus de l'entreprise

- Déployer la CRM
- Compléter le marketing produit avec le marketing client

La Data Science et le Business

Mohamed Heny SELMI



Mohamed Heny SELMI

Domaines d'application

Marketing

- Churn
- Cross selling
- Up selling
- Prédiction sur la durée de vie des clients

Automatisation

- Conduite automatique
- Drones sans pilotes

Ventes et Achats CRM

- Offre à prix réduit
- Prévision sur les demandes

Social media

- Marketing numérique
- Analyse sentimentale

Transport et Logistique

- Tarification dynamique
- Prédiction des retards de vols

Santé publique

- Prédiction des maladies
- Efficacité des traitements
- Segmentation des patients

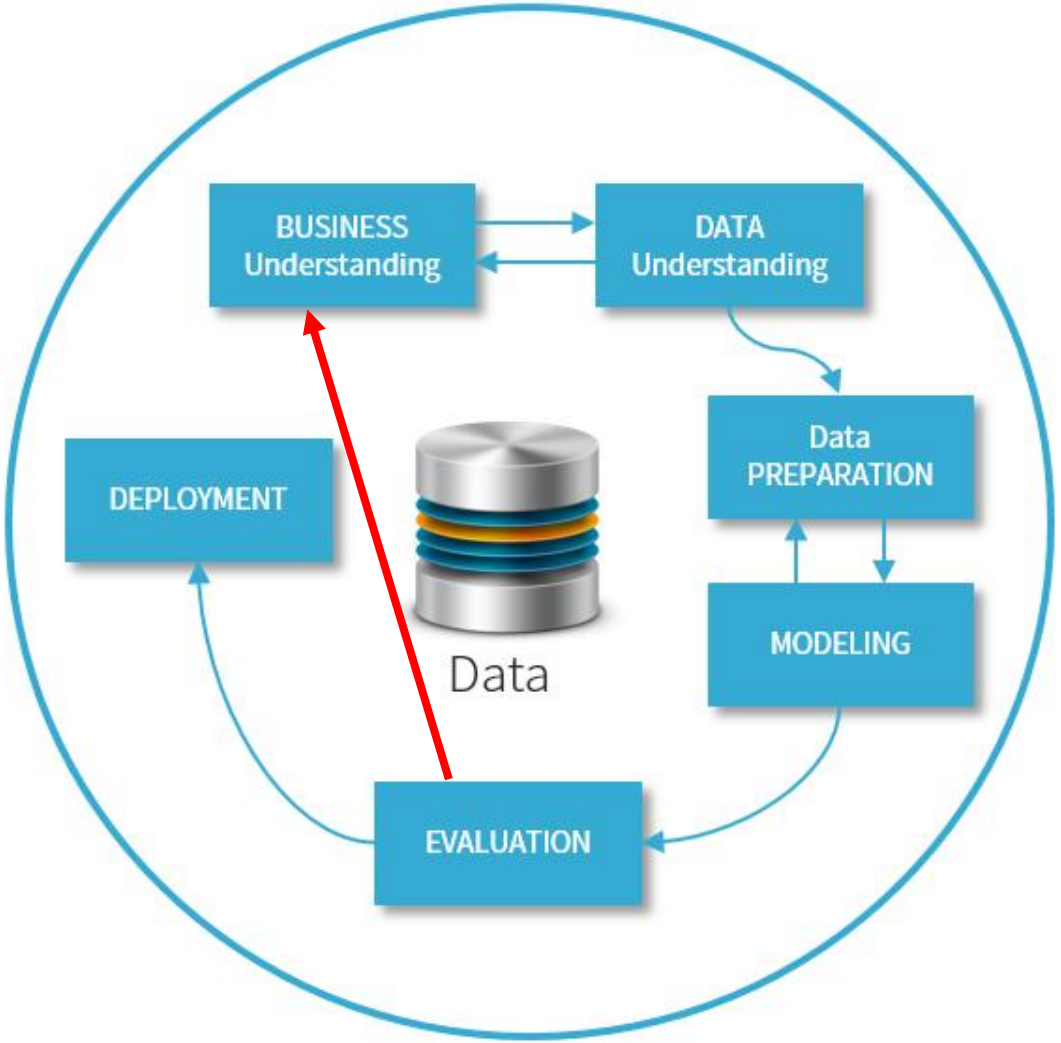
Assurance

- Analyse prédictive des réclamatio
- Détection des fraudes
- Gestion de risque

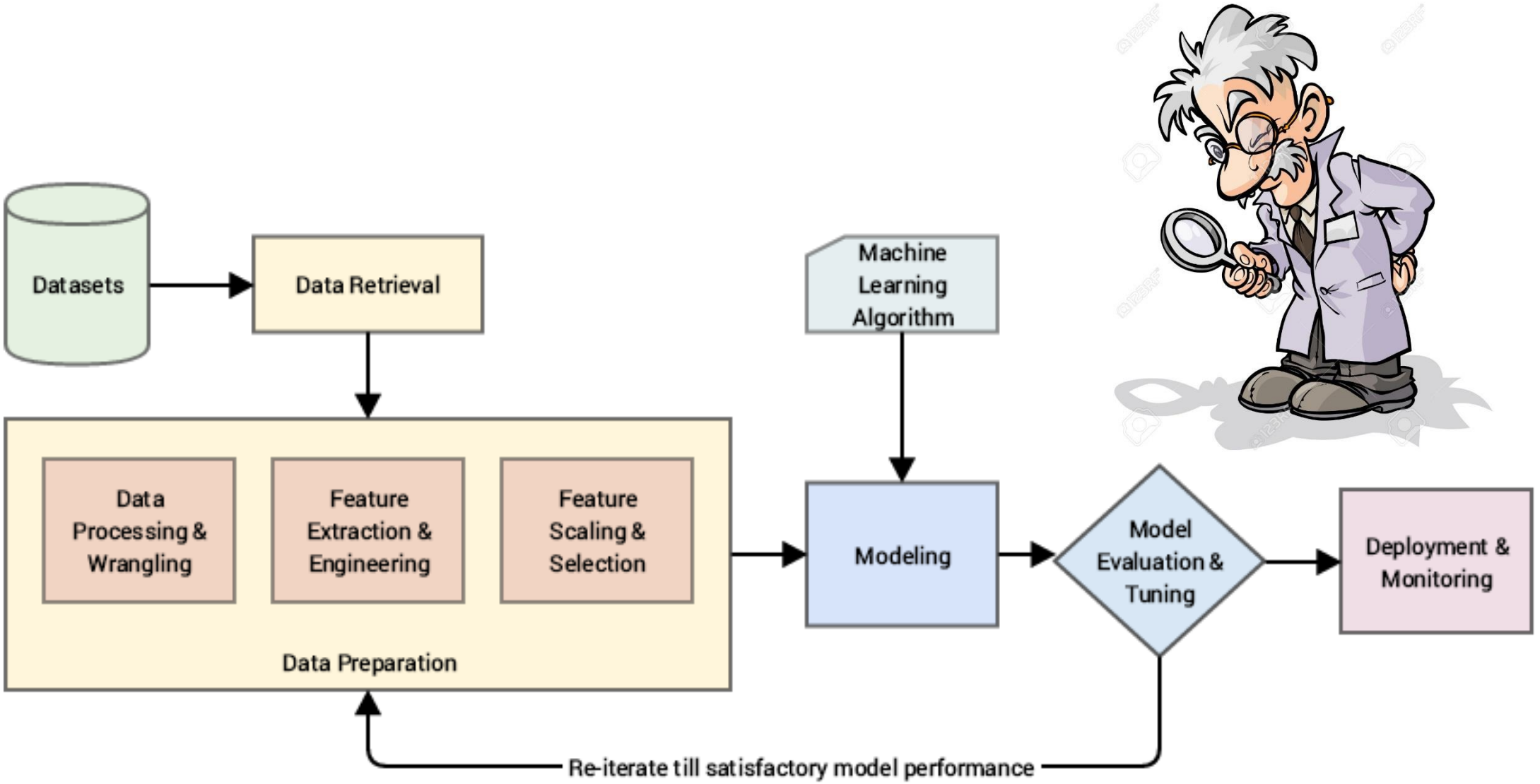
La méthode CRISP (initialement connue comme CRISP-DM) a été au départ développée par IBM dans les années 60 pour réaliser les projets Data Mining.
Elle est toujours la seule méthode utilisable efficacement pour tous les projets Data Science.

CRISP

1. Compréhension du métier
2. Compréhension des données
3. Préparation des données
4. Modélisation
5. Evaluation
6. Déploiement



La méthode CRISP a été officiellement adoptée par *Business & Decision* et son utilisation constitue donc un facteur déterminant à la réussite des projets Data Science.



La méthodologie

- ✓ **La data Science est une démarche**
  - Il faut la décomposer en plusieurs projets ou partie de projets.
  - Il faut différentier itération CRISP et USE CASE.
  - Une itération doit générer plusieurs USE CASE.
- ✓ **L'utilisation de la méthode CRISP est indispensable**
  - L'ensemble des équipes doit suivre la méthode
  - Il n'y a pas une équipe qui comprend le métier, l'autre qui comprend les données, etc.
- ✓ **Seules les deux premières étapes décident à propos de la durée globale du projet**
  - Estimer la durée de chaque itération et de chaque use case
  - La première itération doit être sous la forme d'un POC
- ✓ **La phase de conception de modèle et la phase de mise en production doivent être séparées**
- ✓ **Penser «long terme» sur la gouvernance des données**

Compréhension du Métier

Formuler une problématique : poser des questions ?

Mohamed Heny SELMI

CRISP

Compréhension Métier

- ✓ Note de cadrage
- ✓ bien comprendre les éléments métiers et problématiques que la Data Science vise à résoudre ou à améliorer.
- ✓ des profils fonctionnels et techniques capables de localiser les données à l'intérieur, la périphérie et à l'extérieur de l'organisation.

Les enjeux humains et organisationnels

- ✓ **Prendre en compte tous les aspects organisationnels dans un projet**
- ✓ **Collaboration entre toutes les équipes**
- ✓ **La vraie compétence est encore relativement rare, et ça va durer**
  - Le fait d'avoir des ressources qui programment en R et en Python ne suffit pas !
  - La connaissance des lois statistiques est un prérequis, même en machine Learning
- ✓ **Il faut construire des équipes complémentaires et soudées**
- ✓ **Les principales barrières sont psychologiques**
- ✓ **Il faut être en capacité de montrer les résultats :**

Utiliser de la DataViz et du Data Story Telling

Le dialogue avec les métiers

- ✓ **Il faut comprendre le métier que l'on veut modéliser**
  - Peut nécessiter une immersion complète pendant plusieurs semaines
  - Il faut prévoir plusieurs ateliers avec les métiers
  - L'ensemble des équipes Data Science doit avoir au moins une base sur la compréhension métier

- ✓ **Les enjeux métiers doivent être clairs et soigneusement définis**
  - On pense souvent à ce que ça va apporter, mais…
  - On oublie aussi combien ça va couter de ne pas le faire
  - Maitriser la donnée de son marché veut dire maitriser son marché

- ✓ **Le projet Data Science doit avoir un objectif général**
  - L'objectif pourra se préciser au fur et à mesure du projet

- ✓ **Les métiers connaissent leurs données, et les USE CASE qui vont apporter de la valeur**

- Les ateliers métier permettent de les définir à condition de connaitre les données disponibles

Détermination des objectifs Métier

- Établir un background sur la situation actuelle de l'entreprise.
  - Mieux connaitre les facteur en jeu : les ressources personnel et matériel disponibles, les problèmes, les objectifs.
  - Etudier la structure organisationnelle.
  - Décrire le contexte de la problématique métier.
  - Effectuer une étude de l'existant. (solution existante)
- Chercher toutes les informations possibles sur les objectifs métier.
  - Expliciter le problème à résoudre à l'aide de la Data Science.
  - Enoncer toutes les questions *business* le plus précisément possible.
- Définir l'ensemble des critères de réussite de votre business : objectifs et subjectifs.
  - Fournir une documentation sur les clefs de performance de votre business.
  - Faire de sorte que chaque objectif métier soit lié à un critère de réussite.
  - Indiquer les arbitres décidant si vos critères subjectifs seront atteints, et noter leurs attentes.

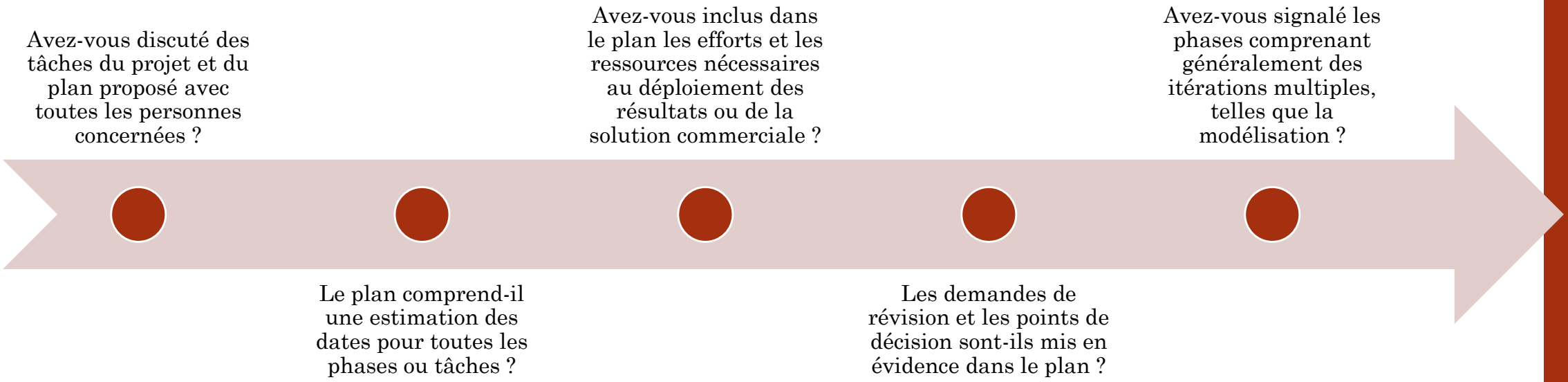
Mohamed Heny SELMI

Détermination des objectifs « Data Science »

- ✓ Décrire la(les) typologie(s) du problème Data Science
- ✓ Fournir des informations concises sur les objectifs Data Science en employant des unités de temps précises.
- ✓ Dans la mesure du possible, exprimer vos résultats désirés sous forme de chiffres.
- ✓ Déclarer les modalités d'évaluations à utiliser.
- ✓ NB : ne pas confondre l'évaluation au sens Data Science et l'évaluation au sens Business

Mohamed Heny SELMI

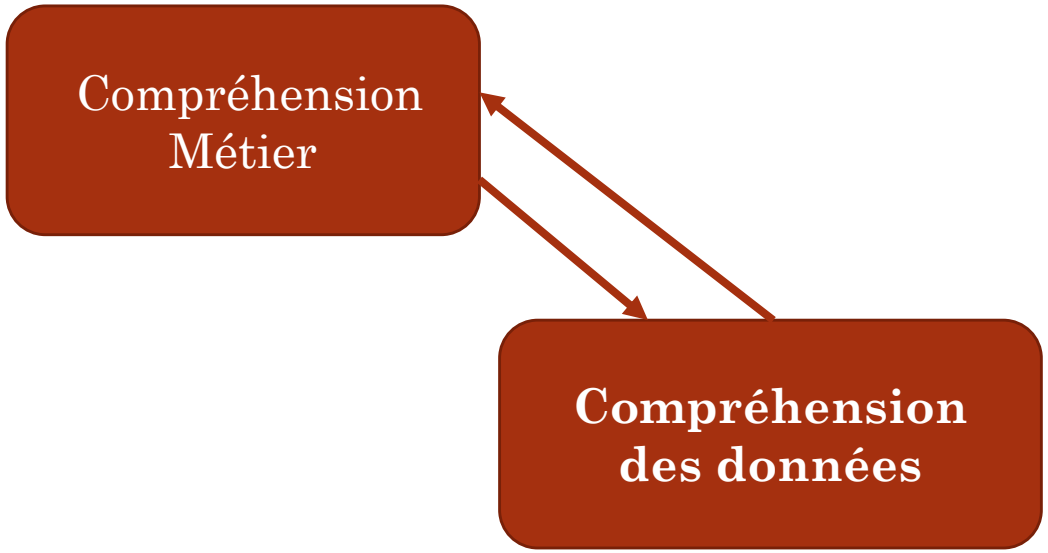
Production d’un plan de projet



Compréhension des données

Mohamed Heny SELMI

CRISP



- ✓ Validation des choix des méthodes statistiques avec un Data Scientist Senior.
- ✓ déterminer précisément les données à analyser, à identifier la qualité des données disponibles et à faire le lien entre les données et leur signification d'un point de vue métier.

Les données

- ✓ **La Data Science est basée sur les données**
- ✓ **On ne peut prédire que ce que l'on peut mesurer**
- ✓ **Les données doivent être disponibles dans leur intégralité**
  - Elles doivent être documentées et structurées
  - Les données non structurées peuvent être converties en données structurées
- ✓ **Les données doivent être riches, complètes et de qualité**
  - On doit utiliser la data science pour améliorer leur qualité
- ✓ **Il ne faut pas se priver d'utiliser systématiquement les données externes en plus des données internes**
  - Utiliser l'open data, les données météorologiques, les données géographiques, etc.
- ✓ **Attention aux données temporelles**

- Des traitements spécifiques : Séries Temporelles…

Context is king !

90 % des données générées :

- **l'ont été dans les deux années 2013 et 2014 !**
- **L'ont été dans l'année 2015 !**

Données internes

Seules 30% des données des entreprises sont exploitées

IoT

Prévision 2020
80 Mds
d'objets
connectés

Open Data

Valeur
estimée de
« l'unloking»
3220 M€

Mobile

63 Mds €
des revenus
d'ici 2018

Social

Prévision 2017
1,7 Mds
d'utilisateurs
fixes
1,4 d'utilisateurs
mobiles

Collecte des données initiales

Données existantes.

Cette catégorie comprend plusieurs types de données, telles que les données transactionnelles, les données issues d'enquêtes, les logs Web, etc. Evaluez ces données existantes pour voir si elles suffisent à répondre à vos besoins.

Données acquises.

Votre société utilise-t-elle des données d'appoint, telles que des données démographiques ?

Si la réponse est non, peut-être faut-il envisager leur utilisation.

Autres données.

Si les sources ci-dessus ne répondent pas à vos besoins, vous devrez peut-être mener des enquêtes ou effectuer davantage de suivis afin de compléter les magasins de données existants.

Collecte et examen des données

Examiner les données

- Quels sont les attributs (colonnes) de la base de données qui semblent les plus prometteurs ?
- Quels sont les attributs qui semblent sans intérêt et peuvent être exclus ?
- Le nombre de données permet-il de tirer des conclusions pouvant être généralisées ou d'effectuer des prévisions précises ?
- Les attributs sont-ils trop nombreux pour la méthode de modélisation choisie ?
- Opérez-vous la fusion de données issues de plusieurs sources ? Si oui, certains points risquent-ils de poser problème lors de la fusion ?
- Avez-vous envisagé le mode de traitement des valeurs manquantes dans chacune de vos sources de données ?

Description des données

Quantité de data

- Les grands ensembles de données peuvent produire des modèles plus précis, mais augmentent également le temps de traitement : utiliser un sous-ensemble de données.
- Lors de la prise de notes en vue du rapport final, veillez à inclure des statistiques sur la taille de tous les ensembles de données

Types de valeur

- Les données peuvent se présenter sous plusieurs formats, tels que le format **numérique**, **catégoriel** ou **booléen** (true/false (vrai/faux)).
- La prise en compte du type de valeur permet d'éviter certains problèmes durant la modélisation.

Méthodes de codage

- Les valeurs d'une base de données représentent souvent des caractéristiques, telles que le sexe ou le type de produit.
  - Par exemple, un ensemble de données peut utiliser les lettres *M* et *F* pour signifier *masculin* et *féminin*, tandis qu'un autre emploiera les valeurs numériques *1* et *2*.
- Notez tous les conflits entre les méthodes de codage dans le rapport sur les données.

Mohamed Heny SELMI

Exploration des données

- **Explorer les données à l'aide des tableaux, des graphiques et des autres outils de DataViz.**
- Formuler des hypothèses
- Elaborer les tâches de transformation des données réalisées durant la préparation des données.

Mohamed Heny SELMI

Ne faites pas trop confiance à vos données

Mohamed Heny SELMI

Vérification de la qualité des données

Les **données manquantes** comprennent les valeurs vides ou codées comme une absence de réponse (telles que *$null$*, *?* , ou *999*).

Les **erreurs de données** sont généralement des erreurs typographiques faites lors de la saisie des données.

Les **erreurs de mesure** représentent notamment les données saisies correctement, mais basées sur une méthode de mesure erronée.

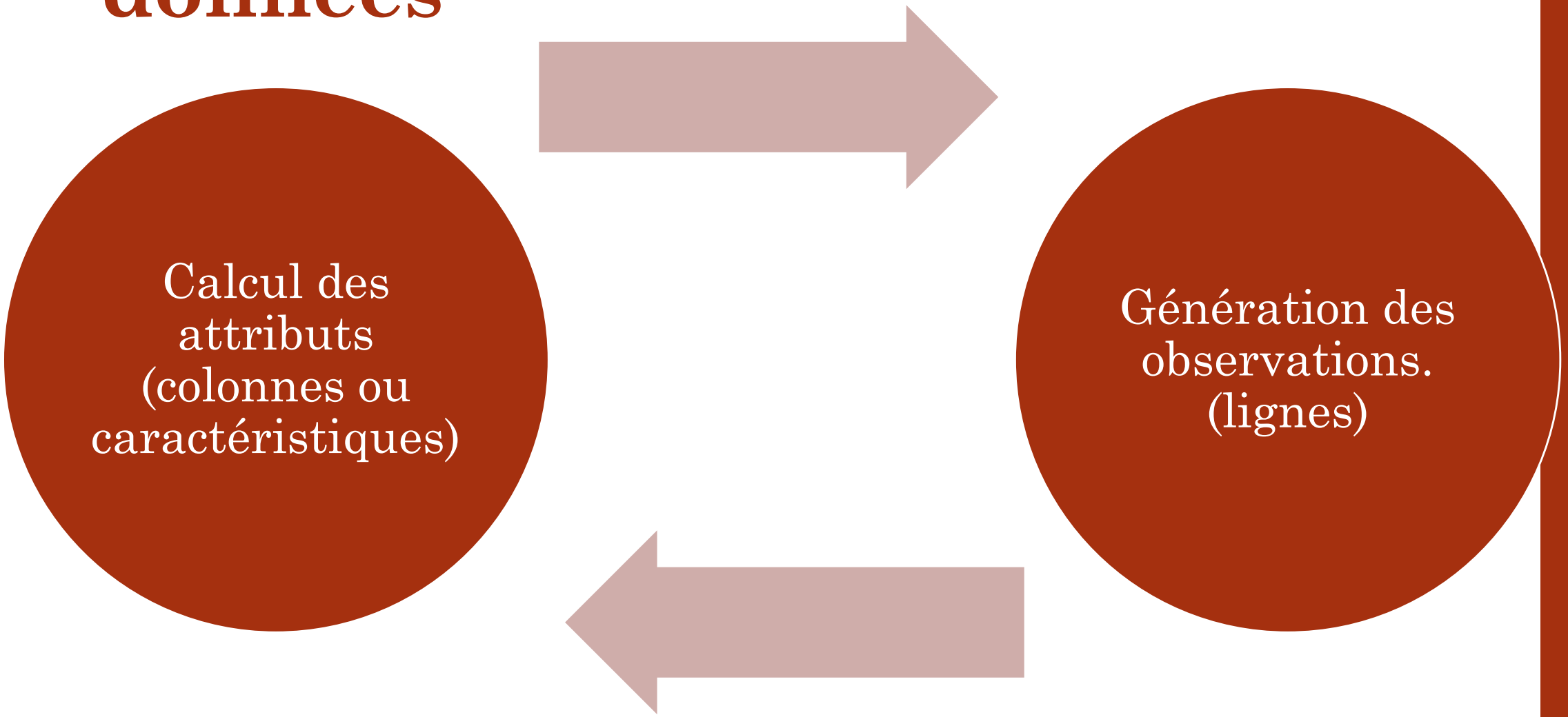
Les **incohérences de codage** concernent généralement les unités de mesure non standard ou les incohérences dans les valeurs, telles que l'utilisation de *M* et de *masculin* pour le sexe.

Les **métadonnées erronées** représentent notamment les discordances entre la signification apparente d'un champ et celle énoncée dans le nom ou la définition du champ.

Préparation des données

Mohamed Heny SELMI

Construction de nouvelles données



Intégration de données

La fusion de données

- qui implique la fusion de deux ensembles de données possédant des observations semblables mais des attributs différents.
- Ces données sont fusionnées à l'aide d'un même identificateur-clé pour chaque observation(tel que l'ID client).
- Les données qui en résultent ont un plus grand nombre de colonnes ou de caractéristiques.

L'ajout de données

- qui implique l'intégration de plusieurs ensembles de données possédant des attributs semblables mais des observations différents.
- Ces données sont intégrées en fonction d'un champ identique (tel qu'un nom de produit ou une durée de contrat).

Mohamed Heny SELMI

Formatage de données

Tenez compte des questions suivantes lors du formatage des données :

- Quels modèles prévoyez-vous d'utiliser ?
- Ces modèles nécessitent-ils l'application d'un format ou d'un ordre particulier aux données ?

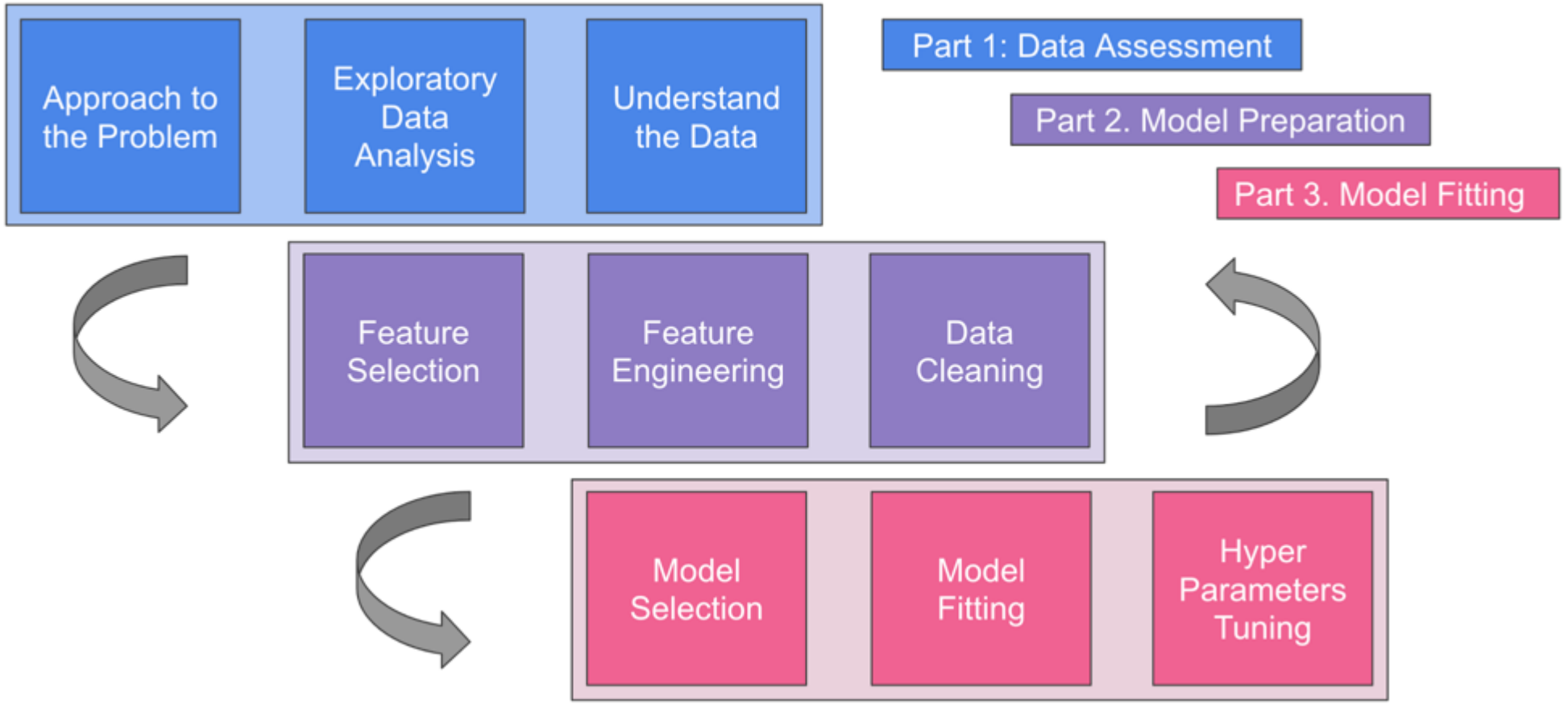
Mohamed Heny SELMI

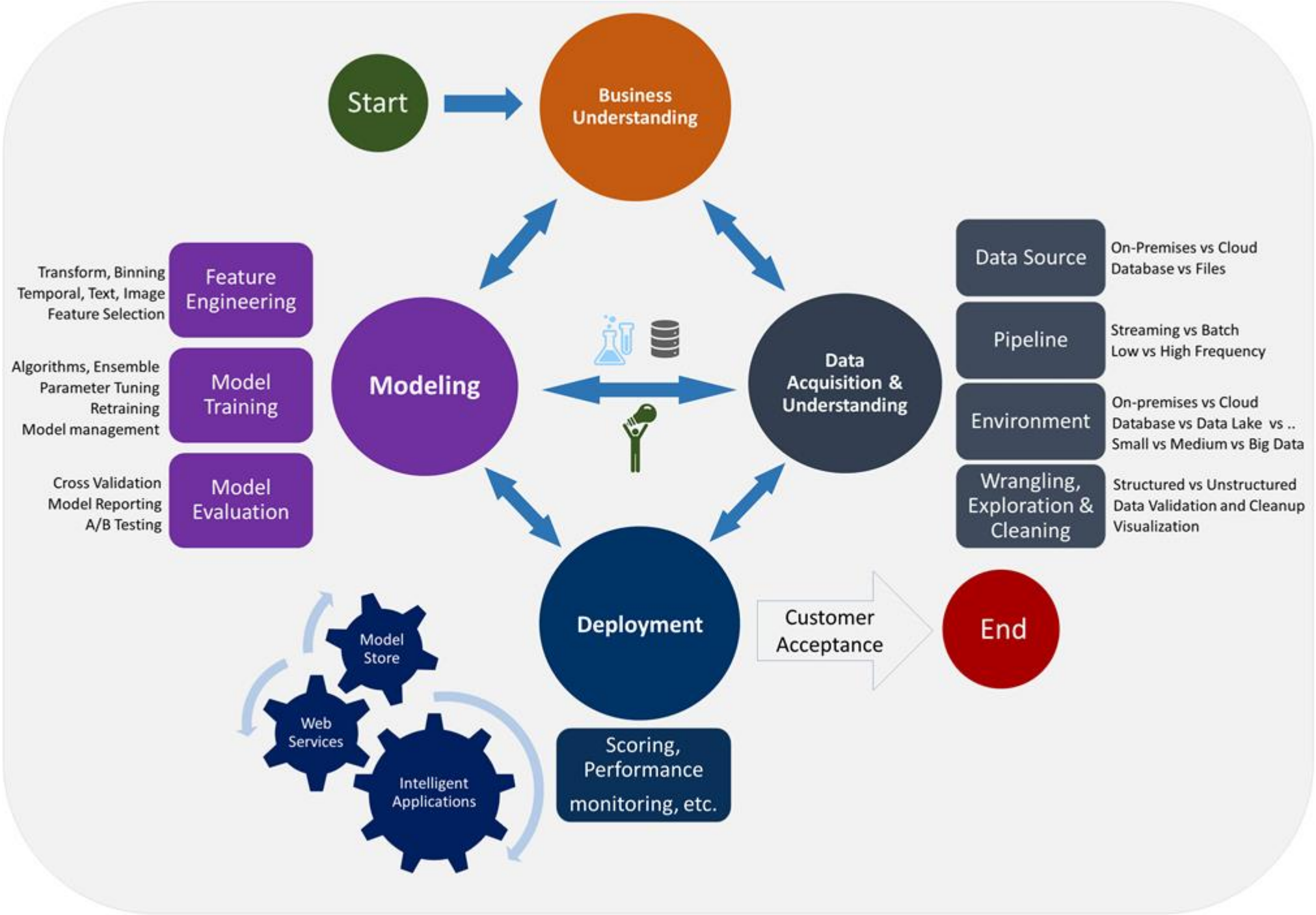
Feature Engineering

Une Définitions…

Le Feature Engineering :

- ✓consiste à utiliser la connaissance du domaine des données pour créer des fonctionnalités qui font alimenter les algorithmes d'apprentissage automatique.
- ✓Si le Feature Engineering est effectuée correctement, ça augmente la puissance prédictive des algorithmes d'apprentissage automatique en créant des fonctions à partir de données brutes qui facilitent le processus d'apprentissage automatique.
- ✓L'ingénierie des fonctionnalités est un art.



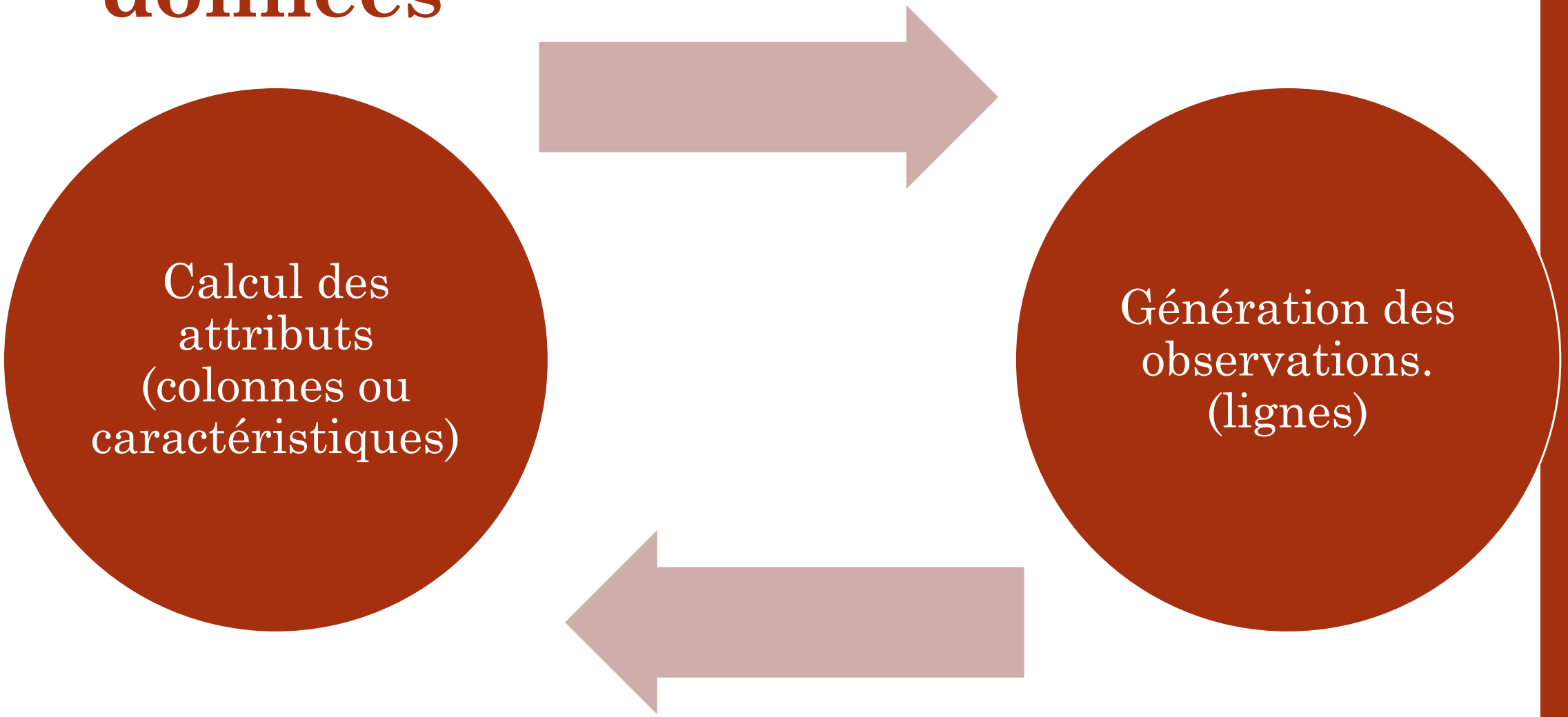


Feature Engineering

Sélection des observations (lignes) : implique des décisions concernant les comptes, les produits ou les clients à inclure.

Sélection des attributs ou des caractéristiques (colonnes) : implique des décisions concernant l'utilisation de caractéristiques telles que le montant des transactions ou le revenu des ménages.

Construction de nouvelles données



Intégration de données

La fusion de données

- qui implique la fusion de deux ensembles de données possédant des observations semblables mais des attributs différents.
- Ces données sont fusionnées à l'aide d'un même identificateur-clé pour chaque observation(tel que l'ID client).
- Les données qui en résultent ont un plus grand nombre de colonnes ou de caractéristiques.

L'ajout de données

- qui implique l'intégration de plusieurs ensembles de données possédant des attributs semblables mais des observations différents.
- Ces données sont intégrées en fonction d'un champ identique (tel qu'un nom de produit ou une durée de contrat).

Mohamed Heny SELMI

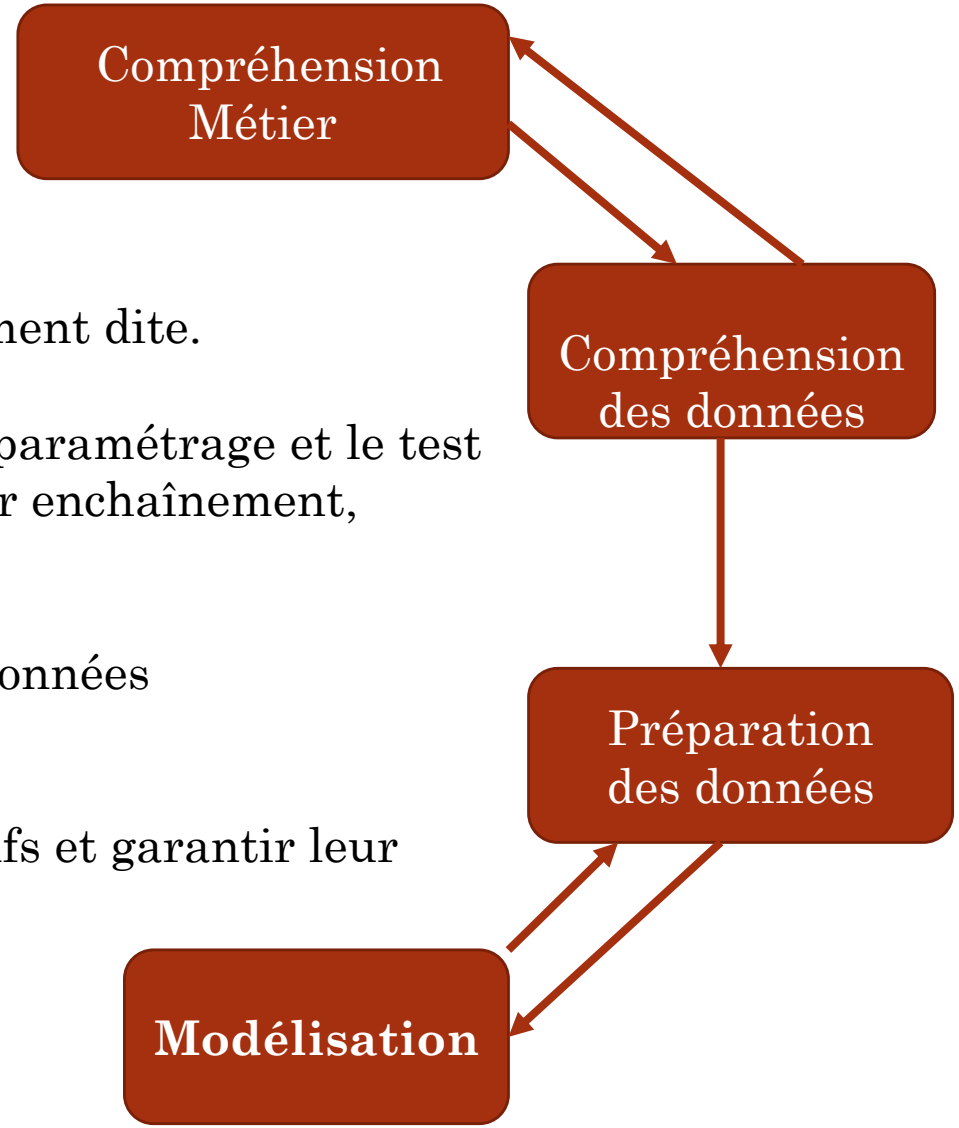
Aspect décisionnel

- ✓ **Nécessité des technologies hybrides**
- ✓ **Stockage lié aux types de données**
- ✓ **Clusterisation versus virtualisation**
- ✓ **Utilisation du Cloud**
- ✓ **Intégration dans le SI**
  - Connecteurs et ETL
  - Intégration avec l'analytique
  - Statistiques et Machine Learning
  - Couplage avec le géo-décisionnel Data

Modélisation

Mohamed Heny SELMI

CRISP



- ✓ C’est la phase de Data Science proprement dite.
- ✓ La modélisation comprend le choix, le paramétrage et le test de différents algorithmes ainsi que leur enchaînement, qui constitue un modèle.
- ✓ Ça consiste à modéliser et croiser les données et donner sens aux corrélations.
- ✓ La mise en œuvre des modèles prédictifs et garantir leur pertinence et leur robustesse.

Les Enjeux Financiers

1

On oublie souvent de combien ça va couter si on le fait pas :

- La data Science est une démarche à moyen terme, il faut des mois pour la mener à bien
- Démarrer en retard, c'est parfois prendre un gros risque de ne pas être prêt à temps face à la compétitivité.

2

Les enjeux financiers ne doivent pas bloquer mais au contraire favoriser la démarche

3

Il faut rapidement construire un «ROAD MAP» dès la fin de la première itération :

- Décomposer itérations et Use Cases
- Elaboration d'une matrice Complexité\Apports
- Prévoir des QUICK WINS

Sélection de techniques de modélisation : Machine Learning

Le choix du modèle le plus adéquat sera généralement basé sur les critères suivants :

- Les types de données disponibles pour l'exploration
- Les *Data Science Goals*.
- Les exigences de modélisations particulières.

Mohamed Heny SELMI

Génération d'une formulation de test

- La description des critères de « qualité d'ajustement »

  d'un modèle

- La définition des données sur lesquelles ces critères

  seront testés

Construction des modèles

- Les **valeurs des paramètres**, qui comprennent les notes que vous avez prises concernant les paramètres aboutissant aux meilleurs résultats.
- Les **modèles** réels produits.
- Les **descriptions des résultats du modèle**, qui incluent les problèmes de performances et de données rencontrés lors de l'exécution du modèle et de l'exploration de ses résultats.

Evaluation des modèles

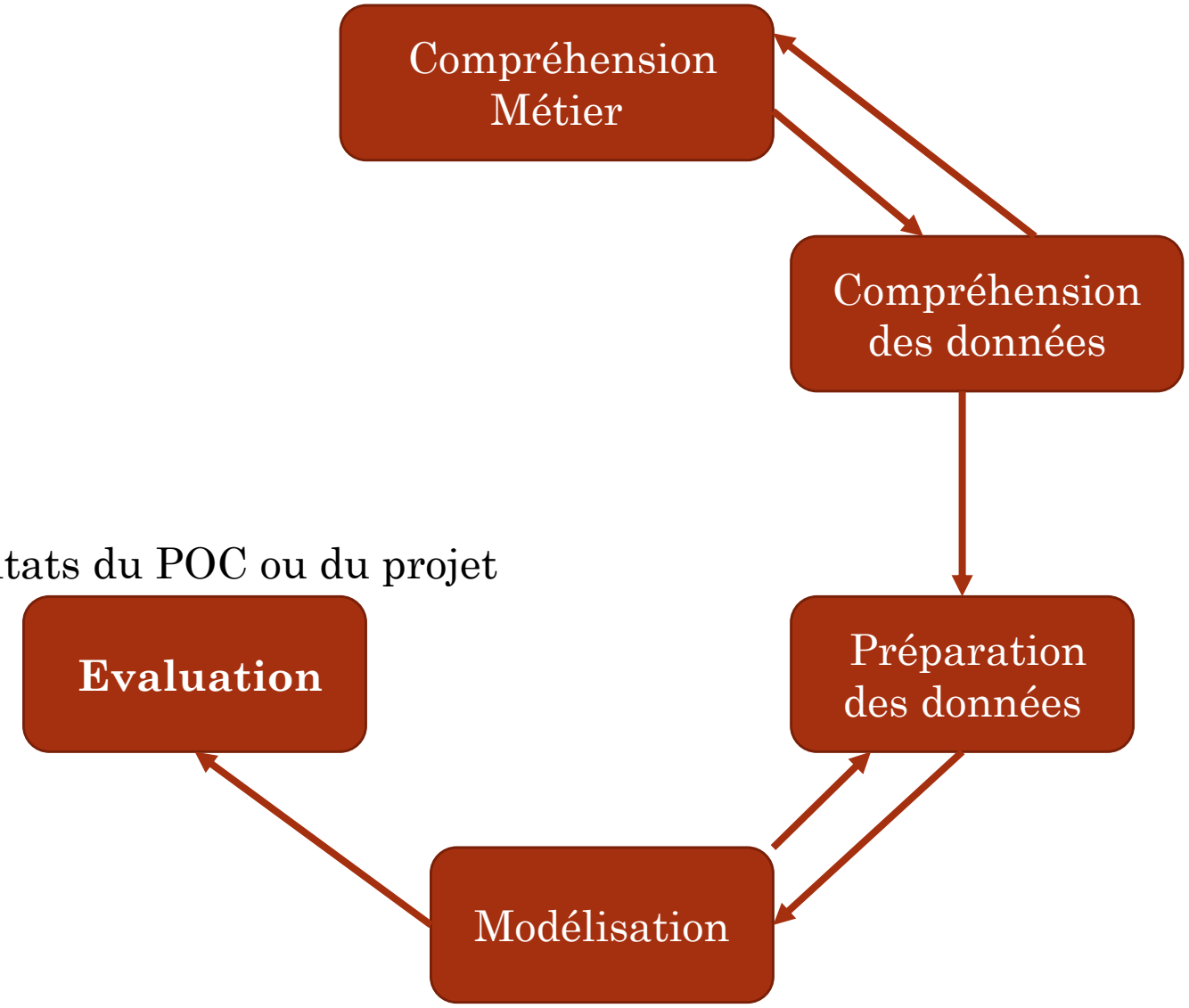
- Analyser les modèles initiaux
- Déterminer ceux qui sont suffisamment précis ou efficaces pour être dit finaux.
- Un modèle final peut désigner un modèle « prêt pour le déploiement » ou un modèle « illustrant des motifs intéressants »

Mohamed Heny SELMI

Evaluation

Mohamed Heny SELMI

CRISP



- Restitution des résultats du POC ou du projet

Les Enjeux Techniques

1. **Le Data Lake ne suffit pas :**
   Nécessité de construire un Data Hub

2. **Attention à l'infrastructure et à l'architecture :**
   - Multiplier les infrastructures ajoute beaucoup de complexité, il vaut mieux éviter.
   - L'infrastructure impacte la méthodologie.

3. **Ne pas partir sur une architecture Big Data si votre projet ne le nécessite pas :**
   - Une architecture Big Data ajoute une charge de travail supplémentaire dans un projet Data Science
   - Big Data n'est pas automatique en Data Science

4. **Attention au choix d'outils et de langages :**
   - Ne pas hésiter à choisir plusieurs outils et langages si besoin
   - R et Python doivent être utilisés mais ne suffisent pas

5. **L'analyse descriptive doit être faite avant toute analyse prédictive**

6. **Il faut utiliser les statistiques et le machine Learning conjointement**
   - Ne jamais faire l'impasse sur l'aspect Statistique du problème à résoudre
   - L'approche statistique permet d'aller jusqu'à un doublement de la précision des modèles par rapport au machine Learning seul

Evaluation des résultats

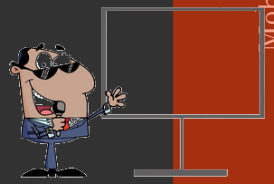
- Vos résultats sont-ils énoncés de façon claire et sous une forme facilement présentable ?
- Existe-t-il des constatations originales ou uniques à mettre en évidence ?
- Pouvez-vous classer les modèles et les constatations par ordre d'applicabilité aux objectifs commerciaux ?
- De façon générale, comment ces résultats répondent-ils aux objectifs commerciaux de votre entreprise ?
- Quelles nouvelles questions vos résultats ont-ils soulevées ? Comment formuleriez-vous ces questions en termes commerciaux ?

Détermination des étapes suivantes

- **Passer à la phase de déploiement.**
  - La phase suivante vous aidera à incorporer les résultats du modèle dans votre processus commercial et à produire un rapport final.
  - Même si vos travaux de Data science ont été vains, vous devez utiliser la phase de déploiement de CRISP pour créer un rapport final à remettre au commanditaire du projet.

- **Revenir en arrière, et affiner ou remplacer vos modèles.**
  - Si vous pensez que vos résultats ne sont pas tout à fait optimaux, envisagez de procéder à une nouvelle modélisation.
  - Vous pouvez utiliser ce que vous avez appris dans cette phase pour affiner les modèles et produire de meilleurs résultats.

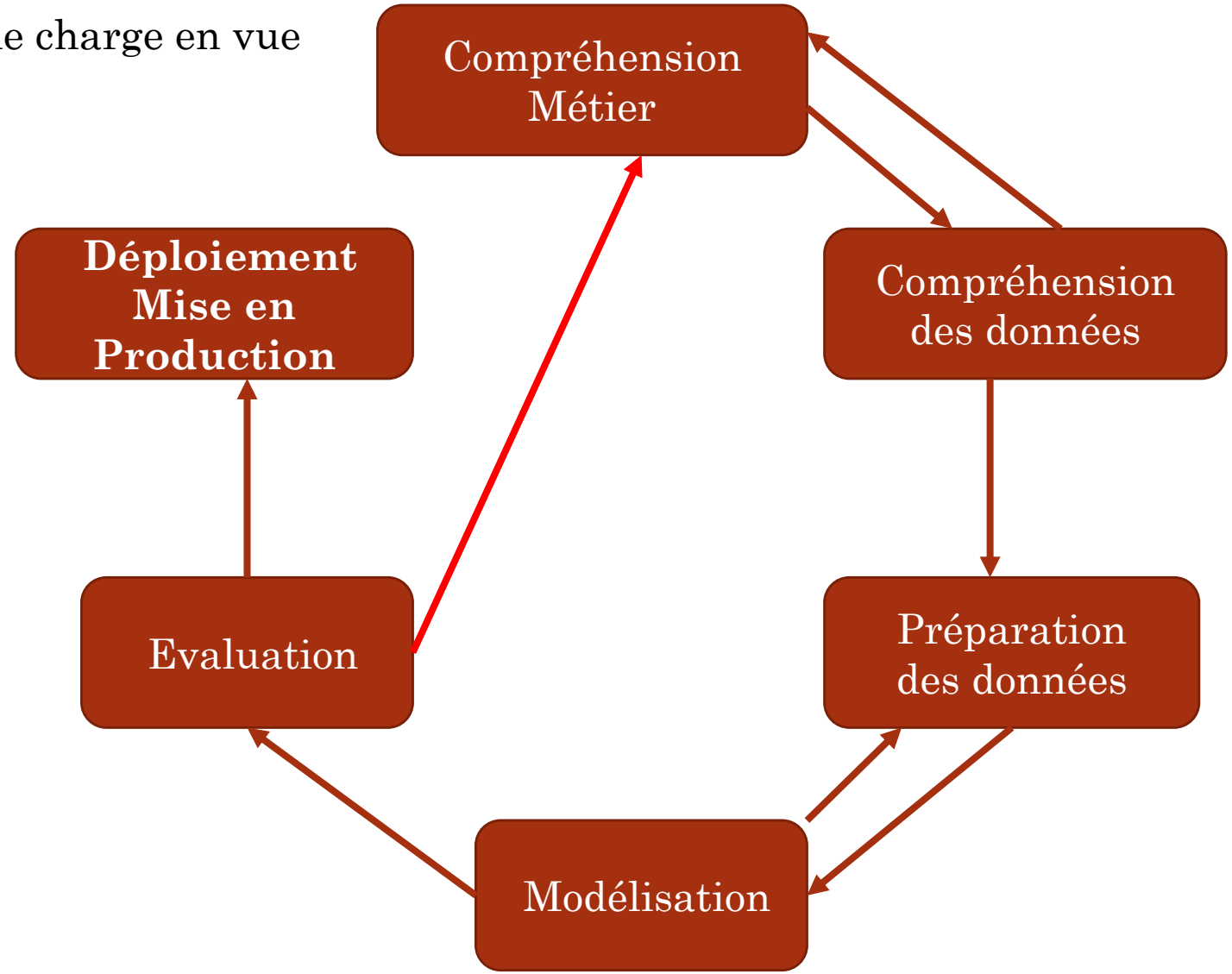
Déploiement



Mohamed Heny SELMI

CRISP

Rédaction du cahier de charge en vue d'industrialisation



Planification du déploiement

- La première étape consiste à récapituler vos résultats (modèles et constatations).
- Vous déterminez ainsi les modèles à intégrer dans vos systèmes de bases de données, ainsi que les constatations à présenter à vos collègues.
- Pour chaque modèle déployable, créez un plan détaillé pour le déploiement et l'intégration dans vos systèmes.
- Notez tout détail technique tel que les exigences de base de données pour la sortie du modèle.

Production d'un rapport final

- Une description complète du problème **Métier** initial
- Le processus utilisé pour effectuer le Data science
- Les coûts du projet
- Des remarques sur tout écart par rapport au plan de projet initial
- Un récapitulatif des résultats de la Data science (modèles et constatations)
- Une présentation du plan proposé pour le déploiement
- Des recommandations pour tout travail de Data science ultérieur, incluant des pistes intéressantes issues de l'exploration et de la modélisation

Mohamed Heny SELMI

MERCI

medheny.selmi@esprit.tn