Machine Learning 4DS

Généralités sur la Data Science

Mohamed Heny SELMI

medheny.selmi@esprit.tn

Enseignant et Responsable option Data Science à ESPRIT

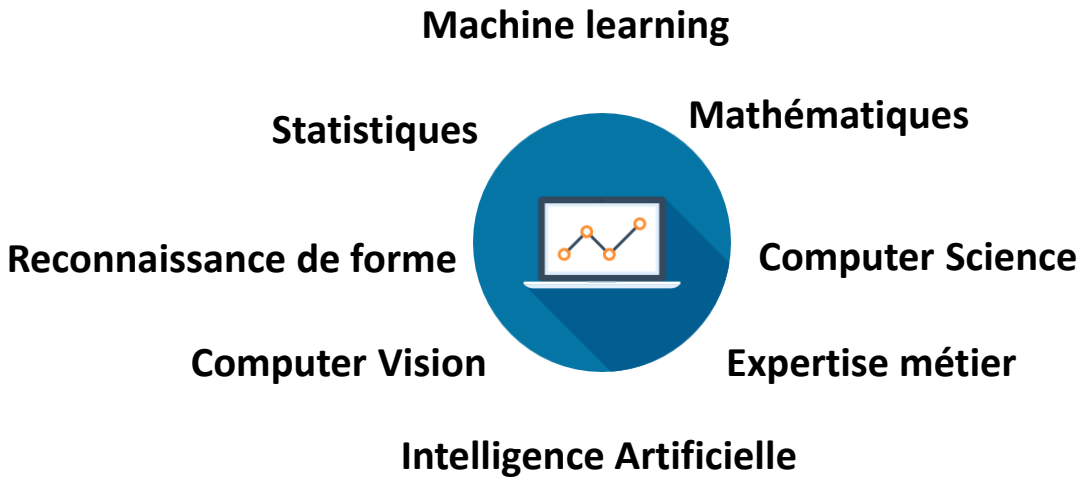
Définitions de la Data Science

"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.

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



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

La Data Science
est une démarche…

KDD / ECD

Knowledge Data Discovery

Extraction de connaissances à partir de données

KDD:
Définition

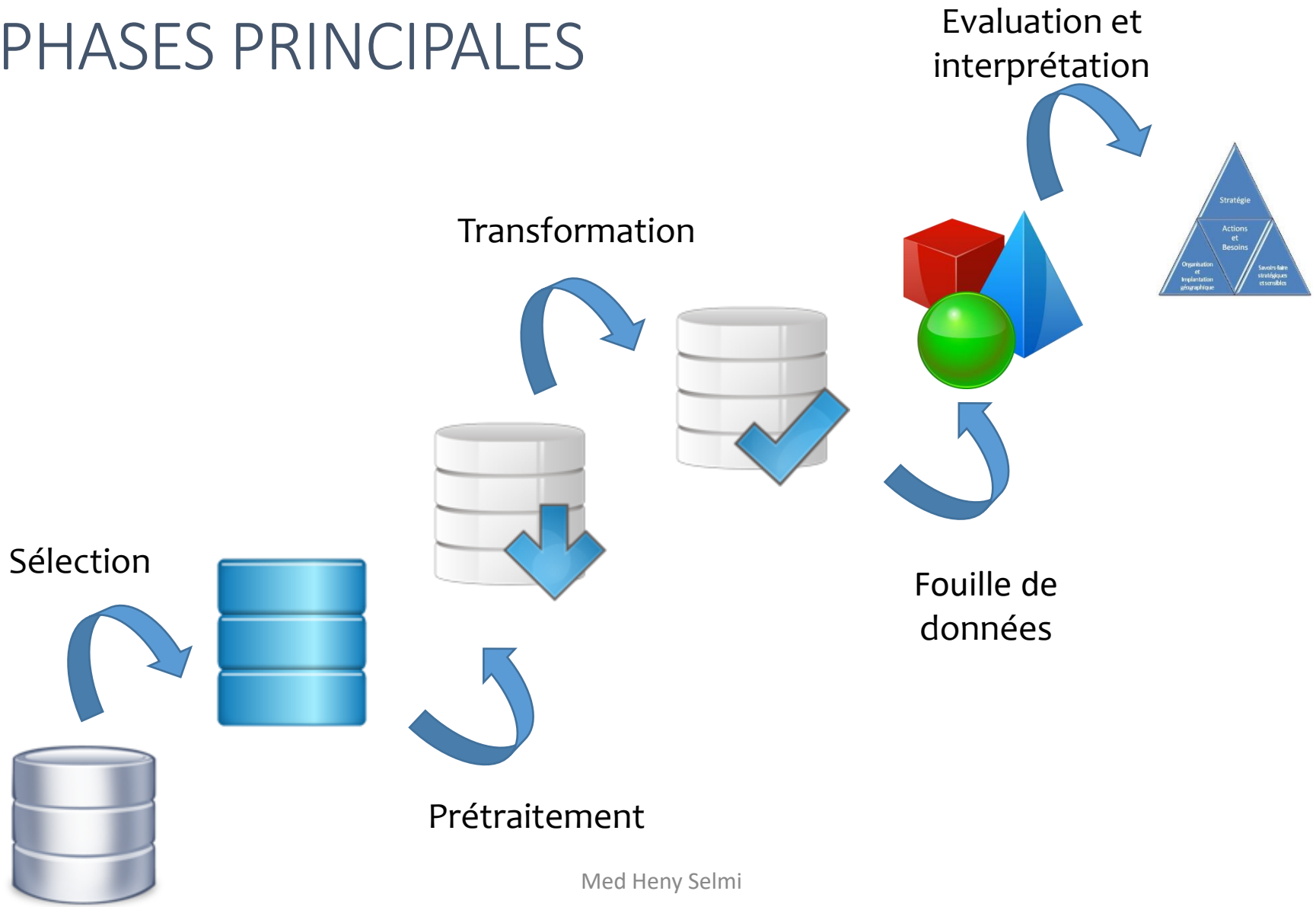
- Knowledge Discovery in Databases

- 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

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 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



CRISP

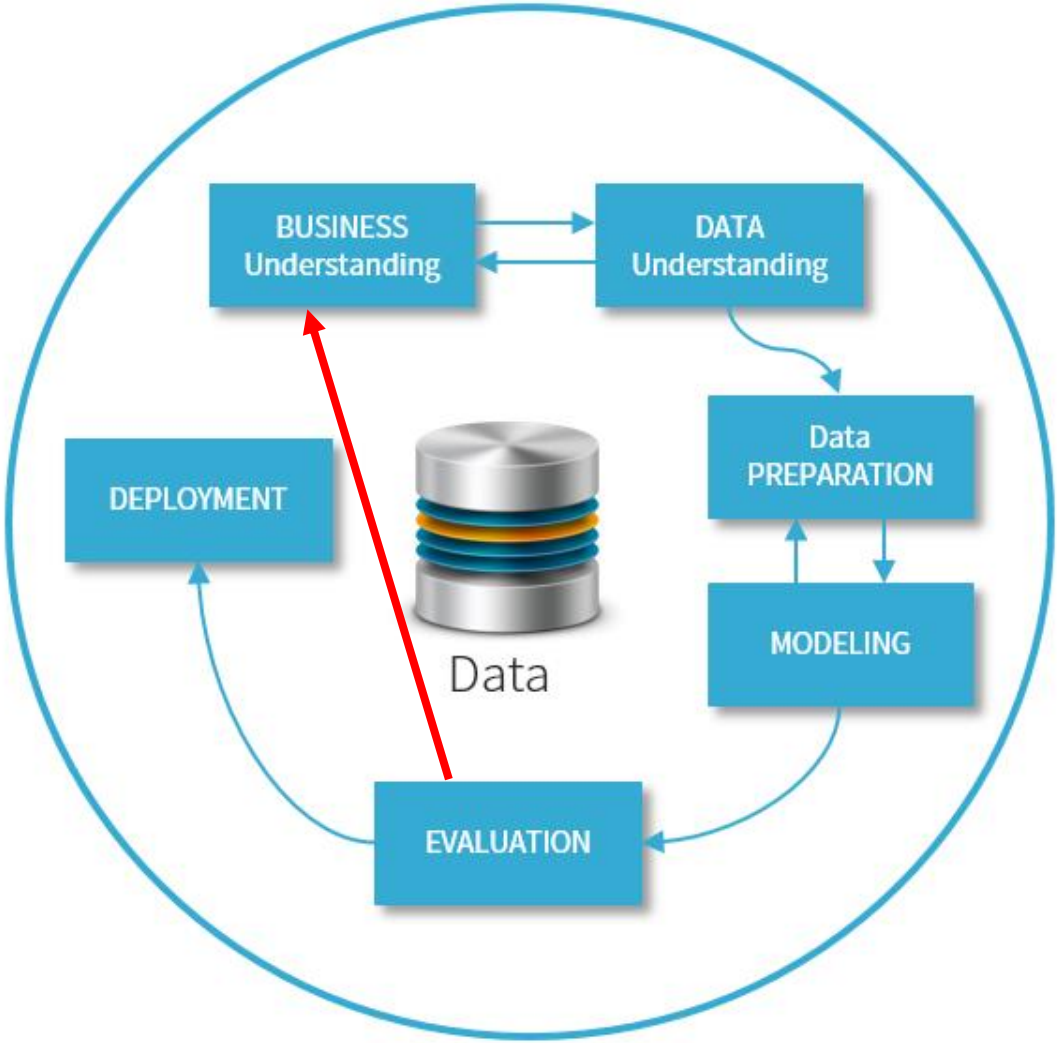
Cross Industry Standard Process

Med Heny Selmi

Méthodologie de travail

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

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.



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.

Compréhension du Métier

Compréhension du problème

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.

Evaluation de la situation

- Quels types de données sont disponibles pour l'analyse ?
  - Quelles sont les sources de données disponibles pour la Data Science ? Notez les types et les formats de données.
  - Comment les données sont-elles stockées ? Pouvez-vous accéder en direct aux entrepôts de données ou aux bases de données opérationnelles ?
  - Projetez-vous d'acquérir des données externes, telles que des informations démographiques ?
  - Existe-t-il des problèmes de sécurité empêchant l'accès aux données requises ?
- Le personnel nécessaire à la réalisation du projet est-il disponible ?
  - Disposez-vous d'experts en matière de données et de commerce ?
  - Avez-vous identifié les administrateurs de base de données ou le personnel d'assistance technique dont vous pourriez avoir besoin ?
- Quels sont les plus grands facteurs de risque en jeu ?
  - Les données ou les résultats du projet sont-ils soumis à des restrictions juridiques ou à des restrictions liées à la sécurité ? *détermination des impératifs*
  - Le déploiement des résultats est-il soumis à des impératifs particuliers (par exemple, la publication sur le Web ou la lecture des scores dans une base de données) ? *détermination des impératifs*
  - Existe-t-il des hypothèses relatives à la qualité des données ? *définition des hypothèses*
  - Disposez-vous de tous les mots de passe nécessaires à l'accès aux données ? *vérification des contraintes*
- Existe-t-il un plan de secours pour chaque risque ?
  - Programmation (Que se passe-t-il si le projet dure plus longtemps que prévu ?)
  - Financement (Que se passe-t-il si le commanditaire du projet rencontre des difficultés budgétaires ?)
  - Données (Que se passe-t-il si les données sont de mauvaise qualité ou peu représentatives ?)
  - Résultats (Que se passe-t-il si les résultats initiaux sont moins spectaculaires que prévu ?)

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

Production d’un plan de projet

- Avez-vous discuté des tâches du projet et du plan proposé avec toutes les personnes concernées ?
- Le plan comprend-il une estimation des dates pour toutes les phases ou tâches ?
- Avez-vous inclus dans le plan les efforts et les ressources nécessaires au déploiement des résultats ou de la solution commerciale ?
- Les demandes de révision et les points de décision sont-ils mis en évidence dans le plan ?
- Avez-vous signalé les phases comprenant généralement des itérations multiples, telles que la modélisation ?

Compréhension des données

Collecte des données initiales[1]

- **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 des données initiales²

Examiner les données et étudier les questions suivantes :

- 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 données.**
  - Les grands ensembles de données peuvent produire des modèles plus précis, mais augmentent également le temps de traitement.
  - Voyez s'il est possible d'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 et à prendre en compte le nombre d'observations et les champs (attributs) lors de la description des 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.

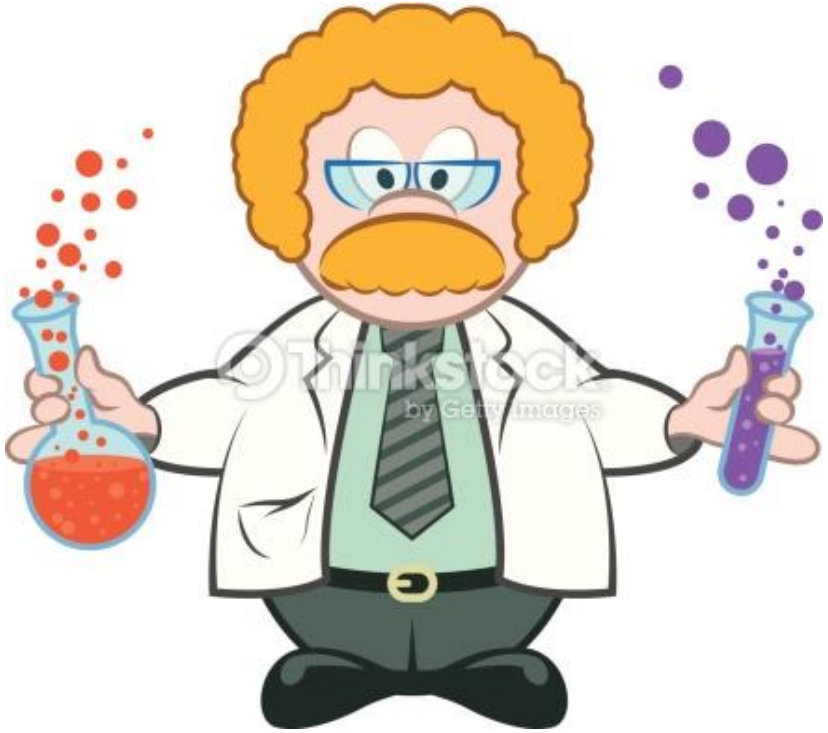
  Par exemple, un ensemble de données peut utiliser les lettres *M* et *F* pour signifier *masculin* et *féminin*, tandis qu'un autre emploiera les valeurs numériques *1* et *2*.
  - Notez tous les conflits entre les méthodes de codage dans le rapport sur les données.

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.

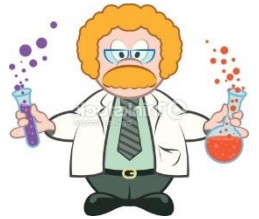
Vérification de la qualité des données

- 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.

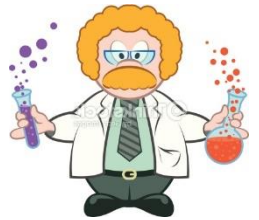


Préparation des données

Sélection de données



- 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.



Nettoyage de données

Problème posé par les données	Solution possible
Données manquantes	Ne rien faire lorsque la portion des NA de l'échantillon est <5% Utiliser un procédure adaptée de remplacement des NA. Excluez les lignes ou les caractéristiques, ou insérez une valeur estimée dans les blancs.
Erreurs dans les données	Procédez de manière logique pour découvrir manuellement les erreurs et les corriger, ou excluez les caractéristiques.
Codage des incohérences	Décidez d'une méthode de codage unique, puis convertissez et remplacez les valeurs.
Métadonnées erronées ou manquantes	Examinez manuellement les champs suspects et recherchez la signification correcte.

Typologie des données manquantes

- **MCAR : Missing Completely at Random** (très rare)
  - ✓La probabilité qu'une valeur de la variable X_1 soit manquante ne dépend pas des valeurs prises par les autres variables $X_{j \neq 1}$, qu'elles soient manquantes ou pas.
  - ✓Il n'est pas possible de définir un profil des individus ayant des valeurs manquantes : la probabilité de ces valeurs est uniforme.
- **MAR : Missing at Random** (peu courant)
  - ✓La probabilité qu'une valeur de la variable X_1 soit manquante ne dépend pas des valeurs prises par les autres variables $X_{j \neq 1}$, mais de leurs valeurs observées.
  - ✓la probabilité d'absence est liée à une ou plusieurs autres variables observées
- **MNAR : Missing not at Random** (le plus fréquent)
  - ✓La donnée est manquante pour une raison précise voulue.
  - ✓La probabilité qu'une valeur de la variable X_1 soit manquante ne dépend pas des valeurs prises par les autres variables $X_{j \neq 1}$, mais de leurs valeurs manquantes.

Méthodes de traitement de DM

- **Exclure les données manquantes**
  - ✓ **List Wise Deletion** : toutes les observations ayant au moins une donnée manquante, cela permet d'effectuer des analyses sur des cas dont toutes les données sont connues.

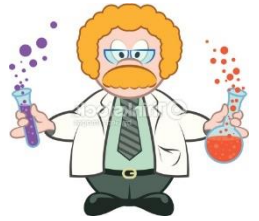
  Peu efficace, car beaucoup d'observations peuvent disparaitre.
  - ✓ **Pair Wise Deletion** : on performe notre analyse avec toutes les cases dont les variables en question sont présentes.

  Point faible : utiliser différentes tailles d'échantillons pour les différentes variables.
- **L'imputation simple**

  Remplacer chaque donnée manquante par une valeur plausible.
  - ✓ **Generalized Imputation** : remplace les valeurs manquantes par la valeur de la moyenne/mediane dans le cas quanti, ou le mode dans le cas quali.
  - ✓ **Similar case Imputation** : remplace les valeurs manquantes par les valeurs provenant d'individus similaires pour lesquels toute l'information a été observée
- **L'imputation multiple**

Procéder à m>1 imputations afin d'obtenir m valeurs pour chaque DM, et à combiner ensuite les stat calculées indépendamment sur m jeux de données.

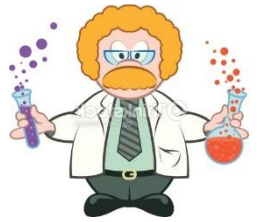
Construction de nouvelles données



- Calcul des attributs (colonnes ou caractéristiques)

- Génération des observations. (lignes)

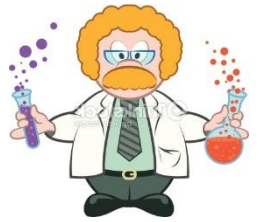
Intégration de données



- 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).

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 ?

Modélisation

Sélection de techniques de modélisation

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.

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 »

Evaluation

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 ?

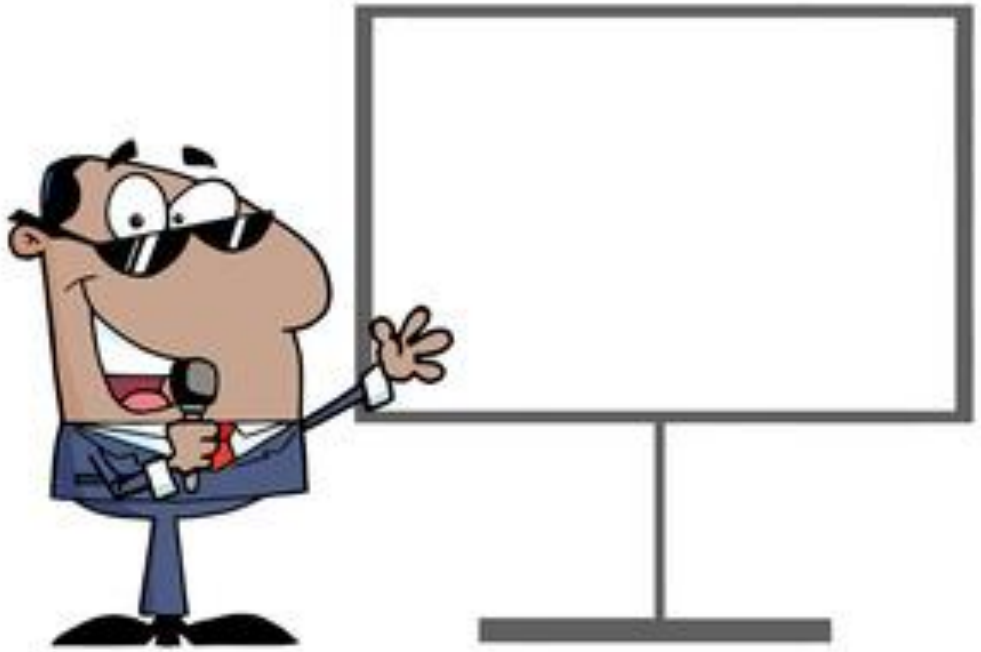
Processus de révision

- Cette étape a-t-elle contribué à la valeur des résultats finaux ?
- Existe-t-il des moyens de simplifier ou d'améliorer cette étape ou opération particulière ?
- Quelles ont été les erreurs ou les échecs rencontrés au cours de cette phase ? Comment peuvent-ils être évités la prochaine fois ?
- Avez-vous constaté que des modèles particuliers ne présentaient aucune perspective d'avenir ? Existe-t-il des moyens de prévoir ces impasses et, par conséquent, de mieux concentrer les efforts ?
- Avez-vous eu des surprises (bonnes ou mauvaises) pendant cette phase ? Avec du recul, existe-t-il un moyen de prédire ces événements ?
- Des décisions ou des stratégies alternatives auraient-elles pu être utilisées lors d'une phase donnée ? Notez ces alternatives en vue de futurs projets de Data science.

Détermination des étapes suivantes

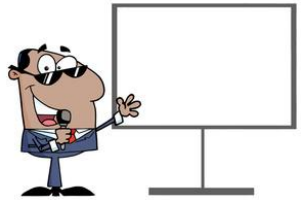
- **Passer à la phase de déploiement.**
  - La phase suivante vous aidera à incorporer les résultats du modèle dans votre processus commercial et à produire un rapport final.
  - Même si vos travaux de Data science ont été vains, vous devez utiliser la phase de déploiement de CRISP pour créer un rapport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

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.

Par exemple, votre système peut exiger que la sortie de modèle soit déployée dans un format délimité par des tabulations.

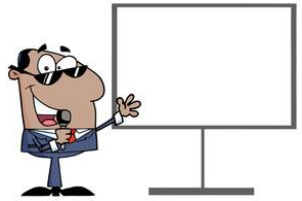
- Pour chaque constatation probante, créez un plan afin de fournir ces informations aux personnes décidant des stratégies.
- Existe-t-il des plans de déploiement alternatifs à mentionner pour les deux types de résultat ?

Par exemple, Sur quels critères déciderez-vous qu'un modèle n'est plus applicable ?

- Prenez en considération le contrôle du déploiement. Identifiez tout problème de déploiement et prévoyez des plans de secours.

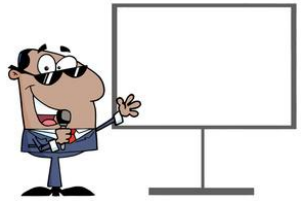
Par exemple, les décideurs peuvent souhaiter obtenir davantage d'informations sur les résultats de la modélisation et plus de détails techniques.

Planification de surveillance/maintenance



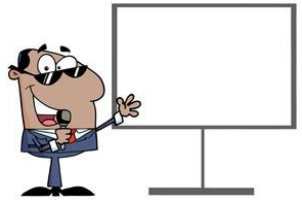
- Pour chaque modèle ou constatation, quels sont les facteurs ou influences (valeur marchande ou variation saisonnière) à prendre en compte ?
- Comment la validité et la précision de chaque modèle peuvent-elles être mesurées et surveillées ?
- Comment déterminerez-vous le moment où un modèle a « expiré » ? Donnez des détails sur les seuils de précision ou sur les changements attendus dans les données, etc.
- Que se passera-t-il lorsqu'un modèle aura expiré ? Pouvez-vous tout simplement recréer le modèle avec des données plus récentes ou apporter de légères rectifications ? Ou les changements seront-ils assez importants pour nécessiter un nouveau projet de Data science ?
- Ce modèle peut-il être utilisé pour des problèmes commerciaux similaires une fois qu'il aura expiré ? Une documentation de bonne qualité se montre ici essentielle pour l'évaluation de l'objectif commercial de chaque projet de Data science.

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

Exécution d'une révision finale du projet



- Vous devez questionner brièvement les personnes impliquées dans le processus de Data Science. Questions à prendre en compte lors des entretiens :
  - Quelles sont vos impressions générales sur le projet ?
  - Qu'avez-vous appris lors du processus (à la fois sur le Data Science en général et sur les données disponibles) ?
  - Quelles sont les parties du projet qui ont fonctionné correctement ?
  - A quel moment des difficultés ont-elles été rencontrées ? Existait-il des informations qui auraient pu éviter la confusion ?
- Une fois les résultats du Data science déployés, vous pouvez également interroger les personnes concernées par les résultats, telles que les clients ou les partenaires commerciaux. Votre objectif est ici de déterminer si le projet était utile et a offert les avantages escomptés.
- Les résultats de ces entretiens, ainsi que vos propres impressions sur le projet, peuvent être récapitulés dans un rapport final ciblé sur les leçons tirées de cette expérience d'exploration des magasins de données.

Les Enjeux Stratégiques

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**

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

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…

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**
  - La data science «dé-silote» les données, à condition que d'autres silos ne fassent pas apparition
- ✓ **Les principales barrières sont psychologiques**
- ✓ **Il faut être en capacité de montrer les résultats**
  - Utiliser de la DataViz et du Data Story Telling
- ✓ **La démarche peut nécessiter une conduite du changement et dans tous les cas un pilotage transverse**

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

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

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

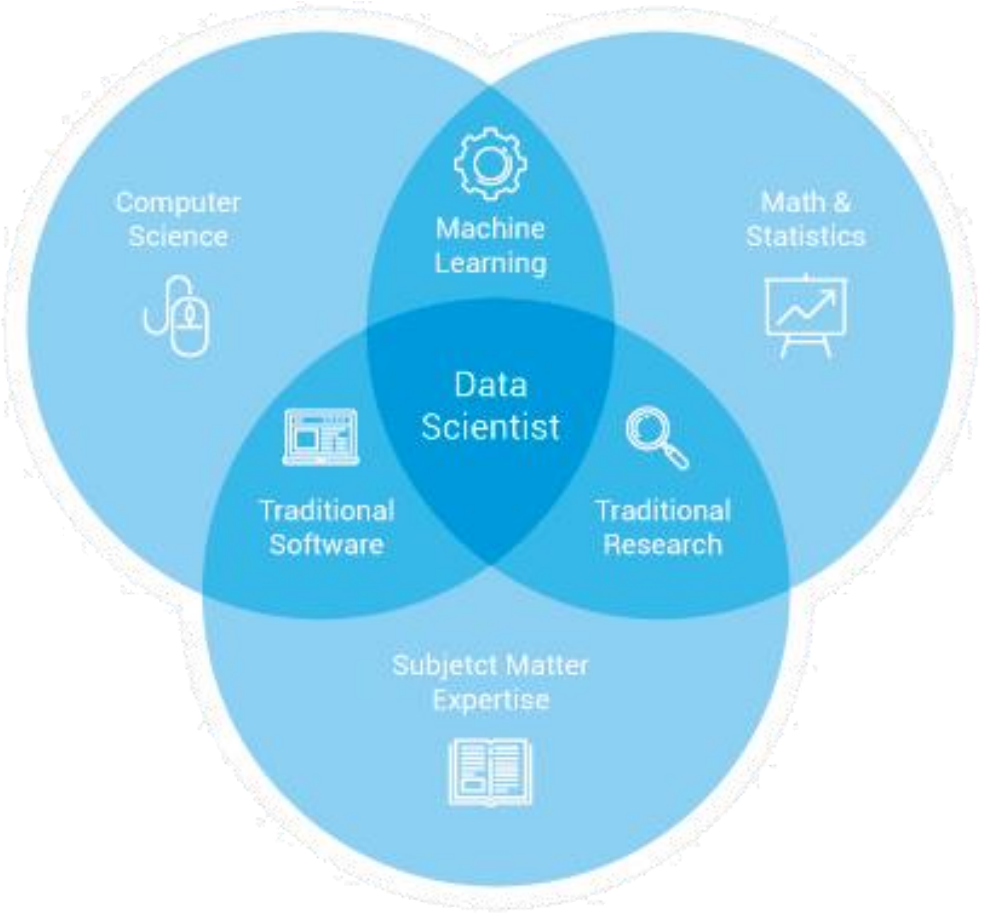
4. **Attention au choix d'outils et de langages :**
   - Ne pas hésiter à choisir plusieurs outils et langages si besoin
   - R et Python doivent être utilisés mais ne suffisent pas

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

Une démarche pluridisciplinaire…

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

Aspect pluridisciplinaire

Informatique, architecture IT

- ✓ Maîtriser un langage de script « haut-niveau », comme Python, R est indispensable dans l'objectif d'être capable de pouvoir développer, tester et valider ses modèles seul.
- ✓ Savoir manipuler un système de gestion de bases de données relationnelles (SQL), ou non relationnelles (technologies NoSQL comme Cassandra ou MongoDB),
- ✓ Connaître les nouvelles technologies de calculs distribués et parallélisés, souvent indispensable pour un traitement efficace de larges volumes de données : MapReduce, écosystème Hadoop (Hive, Pig, etc) et les dernières technologies du Complex Event Processing (Storm), du calcul parallèle ou inmemory (Spark) et de l'indexation (ElasticSearch).
- ✓ Etre au courant des technologies les plus récentes et en cours de développement, mais également des technologies naissantes.
- ✓ Identifier les technologies d'avenir comme Vowpal Wabbit et d'autres librairies d'online learning, libfm pour la factorisation, les packages qui améliorent la performance des algorithmes comme blaze, numba ou xgboost, ou encore les packages qui implémentent les algorithmes ou les méthodes les plus récentes, comme theano pour le Deep learning.
- ✓ Se maintenir au courant des bonnes pratiques et des nouveaux usages des technologies.

Aspect pluridisciplinaire

Mathématiques, statistiques

- ✓ Savoir ce qu'est une distribution, comprendre le principe de l'analyse en composantes principales
- ✓ Connaître les grands tests statistiques
- ✓ Connaître les grands principes : la différence entre l'apprentissage supervisé et non supervisé, les notions d'over-fitting et de validation croisée.
- ✓ Avoir une compréhension générale et le recul nécessaire pour savoir quand et comment appliquer le bon algorithme.

Aspect pluridisciplinaire

Communication, visualisation

- ✓ Expliquer les outils qu'il a développé et les méthodes qu'il a mise en œuvre, en sachant adapter son discours à une audience qui n'a pas forcément un bagage technique avancé.
- ✓ Savoir présenter ses résultats de manière à la fois claire, précise et esthétique est un immense atout.

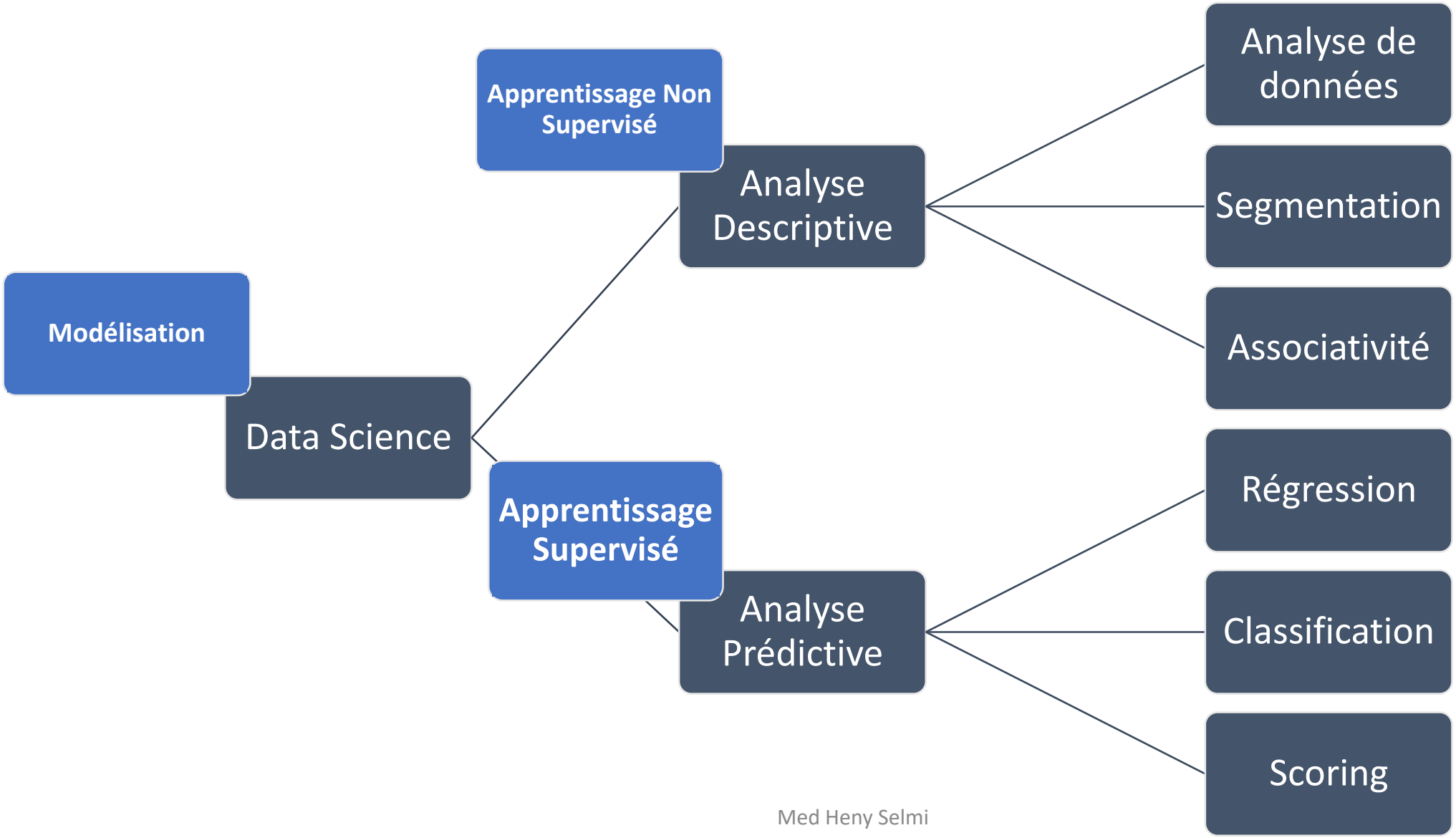
Aspect pluridisciplinaire

Connaissance métier

- ✓ L'objectif du data scientist, est de mettre en œuvre les meilleurs moyens pour répondre à une problématique **métier** précise.
- ✓ Comprendre un enjeu propre à un secteur,
- ✓ formaliser une problématique
- ✓ développer la meilleure approche possible pour y répondre
- ✓ implique une connaissance pointue du secteur et une compréhension en profondeur des données que l'on manipule et une intuition sur les valeurs que l'on peut y trouver.

How do you want that data?

Les méthodes de la modélisation Data Science


Modélisation
Data Science
Apprentissage Non Supervisé
Apprentissage Supervisé
Analyse Descriptive
Analyse Prédictive
Analyse de données
Segmentation
Associativité
Régression
Classification
Scoring