



Fouille de données

▷ Processus & méthodologie

Philippe Lenca

philippe.lenca@telecom-bretagne.eu

Telecom Bretagne
2016-2017



Plan

Méthodologies CRISP-DM-1 CRISP-DM-2 CRISP-DM-3 CRISP-DM-4 CRISP-DM-5 CRISP-DM-6

- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

page 2

Philippe Lenca

Fouille de données ▷ Méthodologies



Plan

Méthodologies CRISP-DM-1 CRISP-DM-2 CRISP-DM-3 CRISP-DM-4 CRISP-DM-5 CRISP-DM-6

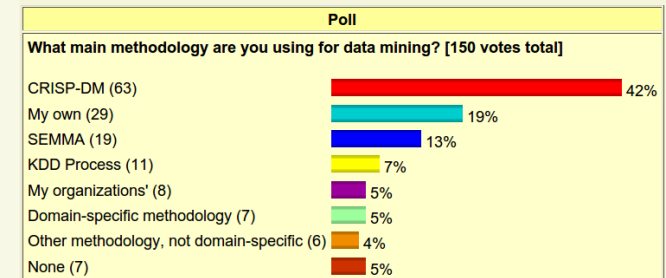
- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

Méthodologies

Méthodologies CRISP-DM-1 CRISP-DM-2 CRISP-DM-3 CRISP-DM-4 CRISP-DM-5 CRISP-DM-6

Un certain nombre ...

[KDNuggets](http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm) : Polls : Data Mining Methodology (Aug 2007)



http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

↔ CRISP-DM, SEMMA et comparaison.

page 4

Philippe Lenca

Fouille de données ▷ Méthodologies



page 3

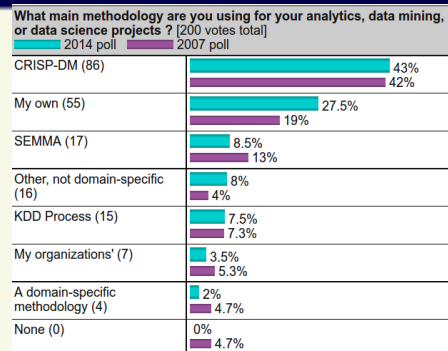
Philippe Lenca

Fouille de données ▷ Méthodologies



Méthodologies

Un certain nombre ...

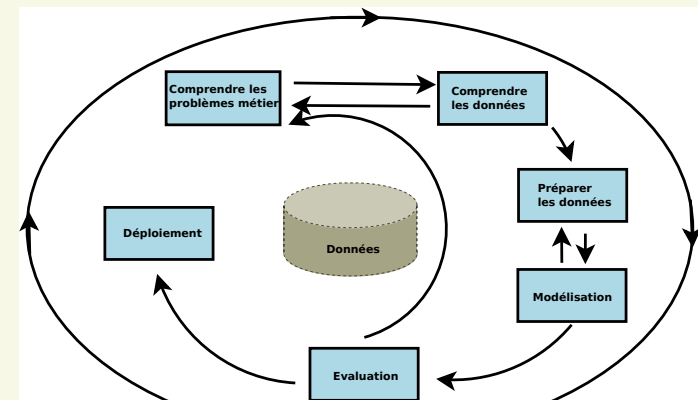


<http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>

CRISP-DM, SEMMA et comparaison.

CRISP-DM

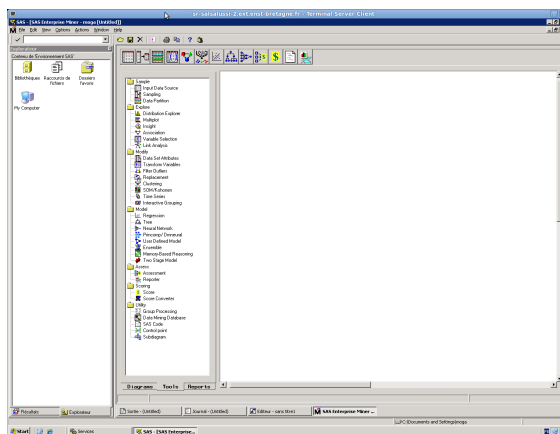
Modèle –cyclique– à six phases



SEMMA

Modèle –cyclique– à 5 phases

- SAMPLE
- EXPLORE
- MODIFY
- MODEL
- ASSESS



SEMMA



CRISP-DM

CRISP-DM-1 : comprendre le(s) problème(s)
 CRISP-DM-2 : comprendre les données
 CRISP-DM-2 : comprendre les données
 CRISP-DM-3 : préparer les données
 CRISP-DM-4 : modélisation
 CRISP-DM-5 : évaluation
 CRISP-DM-6 : déploiement

SEMMA

SAMPLE
 EXPLORE
 MODIFY
 MODEL
 ASSESS

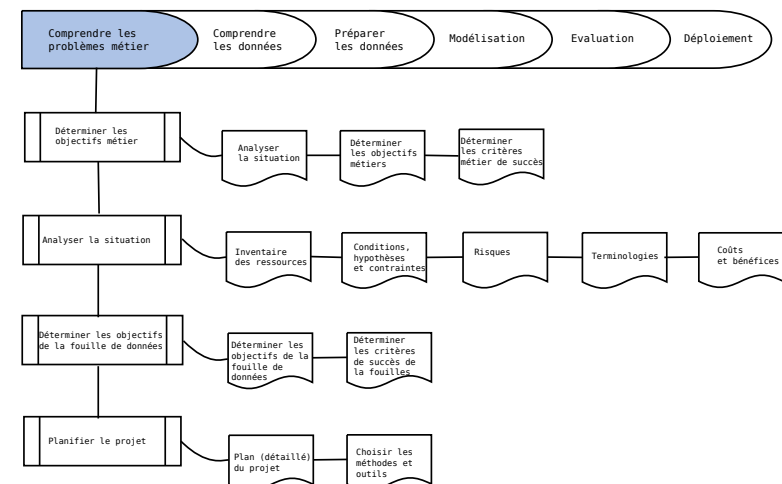
→ Ce cours est fondé sur [Chapman et al., 2001], rapport décrivant la méthodologie CRISP-DM. La lecture de ce rapport est indispensable en complément du cours sur les aspects méthodologiques de l'extraction de connaissances à partir de données.

- 1 Méthodologies en fouille de données
- 2 **CRISP-DM-1 : comprendre le(s) problème(s)**
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

Comprendre le(s) problème(s) métier

- déterminer les objectifs métier
- analyser la situation
- déterminer les objectifs et résultats attendus de la fouille de données
- planifier le projet

→ Etape primordiale, ne pas croire que le problème est évident !



1-1 Déterminer les objectifs métier

- que veut le client ? y a-t-il des objectifs et des contraintes contradictoires ?
- découvrir dès le début les facteurs importants ... qui vont influencer les résultats du travail

↔ Ne pas dépenser d'énergie pour trouver les bonnes réponses aux mauvaises questions

1-2 Analyser la situation

- inventaire **précis de l'existant** : niveau de détail fin par rapport à l'étape précédente sur les
 - ressources
 - contraintes
 - hypothèses

1-1 Déterminer les objectifs métier

Délivrables :

- inventaire de l'**existant** de début du projet
- objectifs métiers **principaux** et questions **relatives**
- **critères de succès** objectifs et subjectifs

1-2 Analyser la situation

Délivrables (**inventaire précis**) :

- des ressources : personnel/expertise, données/accessibilité, capacités de calcul/disponibilité, logiciels/fouille de données/autres
- des conditions, hypothèses et contraintes : planification du projet, compréhension et qualité des résultats, niveau de sécurité des données et des résultats, problèmes de légalité
- des risques : événements pouvant perturber le projet, causes et les actions à entreprendre pour en limiter les impacts
- de la terminologie : glossaire métier et fouille de données
- des coûts et bénéfices : analyse des coûts du projet et des bénéfices potentiels

1-3 Déterminer les objectifs et résultats attendus de la fouille de données

- objectifs métier selon la terminologie métier
- objectifs de la fouille selon la terminologie fouille

1-4 Planifier le projet

- décrire le plan prévu afin d'atteindre les objectifs de la fouille de données et ainsi les objectifs métier
- décrire toutes les étapes de façon détaillée
- y compris une première sélection des outils et techniques

1-3 Déterminer les objectifs et résultats attendus de la fouille de données

Délivrables :

- objectifs : décrire les résultats techniques de la fouille de données permettant d'atteindre les objectifs métiers
- critères de succès : décrire les critères, les mesures de succès techniquement et objectivement, identifier qui évalue subjectivement

1-4 Planifier le projet

Délivrables :

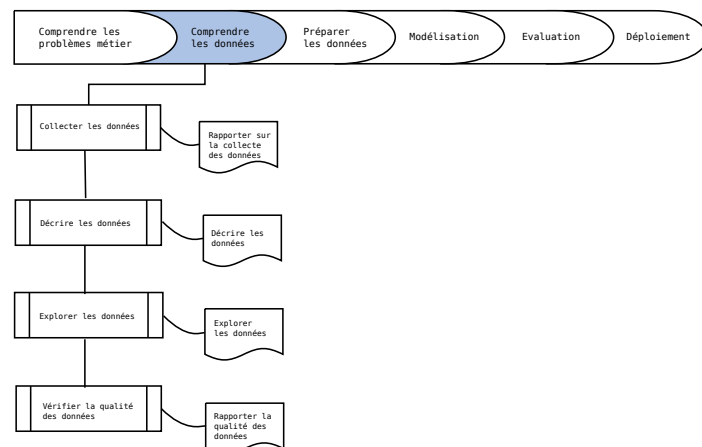
- liste de toutes les étapes avec durée, ressources, entrées, sorties, planification, risques, dépendances
- le document de projet est dynamique, à la fin de chaque étape une adaptation est nécessaire
- une évaluation et un pré-choix des outils

- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 **CRISP-DM-2 : comprendre les données**
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

Comprendre les données

- collecter les données -initiales-
- décrire les données
- explorer les données
- vérifier la qualité des données

↔ Etape permettant de maîtriser les ressources données et leurs caractéristiques.



2-1 Collecter les données -initiales-

- accéder (voire charger) les données listées dans les ressources du projet
- intégrer les données si nécessaire dans le cas de sources multiples
- si nécessaire faire une première étape de préparation des données (pour chargement dans un outil spécifique)

2-1 Collecter les données -initiales-

Délivrables :

- liste des bases de données récupérées
- localisation des bases de données
- méthodes de récupération
- problèmes rencontrés
- solutions apportées

2-2 Décrire les données

Délivrables : rapport décrivant les données

- format
- quantité (individus, attributs)
- attributs

2-2 Décrire les données

- réaliser un examen rapide, de surface, des données

↔ Est-ce que les données satisfont les conditions nécessaires ?

2-3 Explorer les données

- réaliser des analyses sur les données (statistique descriptive, requêtes, visualisation, reporting)
- distributions des valeurs
- relation entre attributs
- propriétés au sein de sous-populations

↔ Répondre à certains buts du projet. Affiner la description des données et le rapport sur la qualité, alimenter l'étape de préparation des données.

2-3 Explorer les données

Délivrables : rapport d'exploration des données

- caractéristiques des données
- hypothèses initiales et impact sur le déroulement du projet
- données quantitatives, graphiques, schémas

2-4 Vérifier la qualité des données

- complétude (couverture des situations nécessaires)
- exactitude (y a-t-il des erreurs, des incertitudes, en quelle quantité ?)
- valeurs manquantes (si oui, comment sont-elles représentées, où apparaissent-elles, en quelle quantité ?)

↔ Répondre à certains buts du projet. Affiner la description des données avec un rapport sur la qualité, alimenter l'étape de préparation des données.

2-4 Vérifier la qualité des données

Délivrable : rapport de qualité des données

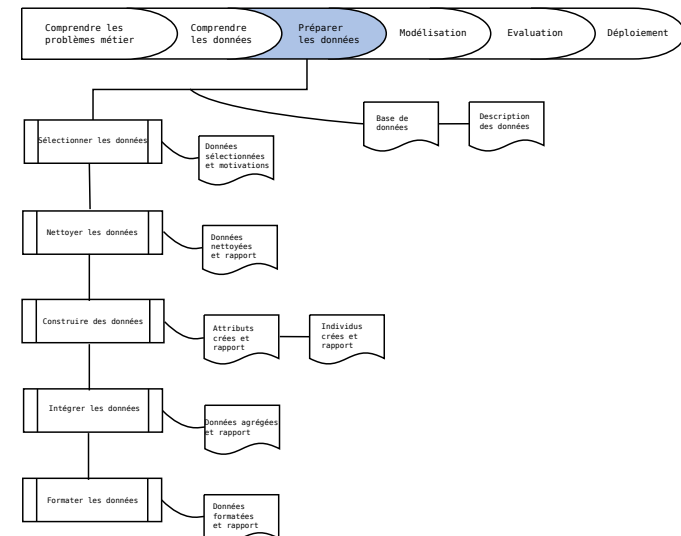
- liste des résultats de la vérification de la qualité
- des solutions possibles (connaissance des données et du métier)
- des causes de la non qualité, des remèdes

- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

Préparer les données

- sélectionner les données
- nettoyer les données
- construire les données
- intégrer les données
- formater les données

↪ Produire les bases de données et la description des données qui seront utilisées pour les tâches de modélisation et d'analyse.



3-1 Sélectionner les données

- sélectionner les données qui seront utilisées pour l'analyse
- adéquation par rapport aux objectifs métiers et de l'analyse
- adéquation par rapport à la qualité des données
- adéquation par rapport à des contraintes techniques (volumétrie, types, format)

3-1 Sélectionner les données

Délivrable : données sélectionnées et critères

- liste des données sélectionnées/exclues (individus, attributs)
- critères de sélection
- motivations pour la sélection/exclusion

3-2 Nettoyer les données

- élever la qualité des données aux niveaux requis par les techniques d'analyse et les objectifs de résultats
- sélectionner des sous-ensembles de qualité
- accepter, contrôler certains défauts
- corriger des erreurs, estimer des valeurs

3-3 Construire les données

- créer des attributs à partir d'autres attributs
- créer des individus
- transformer des valeurs

3-2 Nettoyer les données

Délivrable : rapport de nettoyage des données

- lister les actions de nettoyage
- problèmes rencontrés
- transformations réalisées
- impact sur les résultats

3-3 Construire les données

Délivrable : rapport de construction de données

- lister les nouveaux attributs, nouveaux individus, valeurs transformées
- règles de création (individus et bases d'origine, mise à jour, méthodes de construction)
- règles de transformation (attributs et échelles d'origine, méthodes de transformation)
- motivations

3-4 Intégrer les données

- intégrer différentes sources (bases de données, individus, etc.)
- créer de nouvelles bases, de nouveaux individus ou de nouvelles valeurs

↔ Combiner des informations de sources diverses.

3-5 Formater les données

- modifier syntaxiquement les données (pas de modification sémantique)
- changer la représentation des données (individus, attributs)

↔ Rendre compatibles les données avec les exigences des outils.

3-4 Intégrer les données

Délivrable : rapport d'intégration de données

- lister les données fusionnées/agrégées
- règles de fusion/agrégation (individus, attributs et bases d'origine, mise à jour, méthodes de fusion)
- réviser si nécessaire l'étape de sélection des données
- motivations

3-5 Formater les données

Délivrables : rapport des modifications syntaxiques, données formatées

- lister les modifications
- ordonnancement des attributs
- tri/mélange des individus
- supprimer/ajouter des caractères
- contraindre la taille des attributs
- motivations

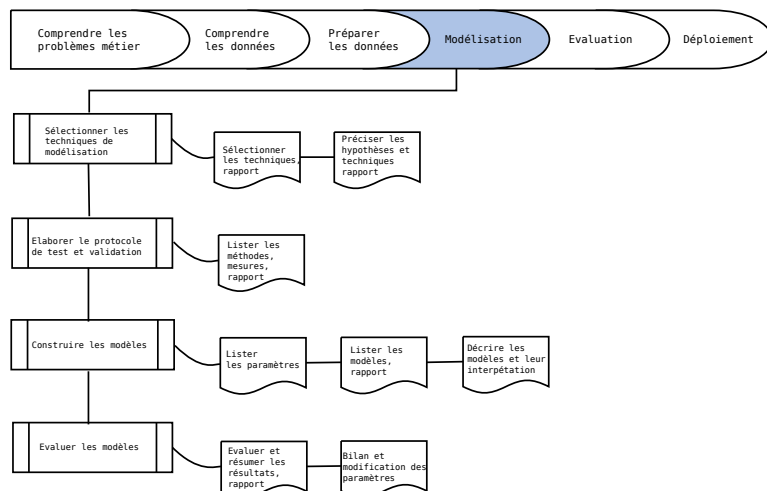
- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation**
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

Modélisation

Choix des méthodes (techniques de modélisation, techniques d'évaluation des résultats, etc.) pour extraire les connaissances :

- sélectionner les techniques de modélisation
- élaborer le protocole de test et validation
- construire les modèles
- évaluer les modèles

↔ Produire des connaissances valides, intéressantes, etc.



4-1 sélectionner les techniques de modélisation

- sélectionner les méthodes de modélisation
 - extraction d'associations
 - classification
 - prédiction
 - etc.
- sélectionner les techniques spécifiques
 - prédiction par arbre de décision vs. réseaux de neurones
 - classification par classification ascendante hiérarchique vs. k-means
 - etc.

↔ Sélectionner les méthodes, les techniques spécifiques permettant de produire des connaissances (disponibilité des outils retenus?).

4-1 sélectionner les techniques de modélisation

Délivrables : rapport sur les techniques utilisées et les hypothèses nécessaires

- documenter les méthodes de modélisation
- documenter les techniques de modélisation
- liste des hypothèses nécessaires sur les données
 - distributions des attributs
 - absence de valeurs manquantes
 - attribut de classe de type symbolique
 - etc.

4-2 élaborer le protocole de test et validation

- définir les méthodes pour tester la qualité et la validité des résultats
- sélectionner les protocoles de test (échantillonnage des données)
- sélectionner ou définir les mesures de qualité

4-2 élaborer le protocole de test et validation

Délivrables : rapport sur le protocole de test et validation

- lister les méthodes
- lister les protocoles d'échantillonnage (bases d'apprentissage, de test et de validation ; autre protocole)
- lister les mesures

4-3 construire les modèles

- exécuter les méthodes retenus à l'aide des outils choisis
- i.e. exécuter des algorithmes instanciés sur des données préparées

⇒ Créer un ou plusieurs modèles à partir des outils et des bases de données préparées.

4-3 construire les modèles

Délivrables : rapport sur les paramètres retenus, les modèles et rapport sur les modèles

- les algorithmes/méthodes sont paramétrisables (les outils offrent plus ou moins de liberté), lister les paramètres et leur valeurs, motivations
- les modèles effectivement générés
- description des modèles, de leur interprétation, des difficultés rencontrées, de leur sens

↔ Documenter les résultats.

4-4 évaluer les modèles

- évaluer techniquement les modèles à partir des critères de succès, des algorithmes, des protocoles de validation (évaluation de l'analyste)
- évaluer les résultats du point de vue métier en faisant appel aux analystes métier, aux experts du domaine

↔ Discuter et documenter les résultats des modèles.

4-4 évaluer les modèles

Délivrables : rapport sur les qualités et défauts des modèles, révisions à apporter

- résumer les résultats
- qualités des modèles
- défaut des modèles
- classement des modèles
- bilan et plan de modification des paramètres

↔ Itérations de construction de modèles.

- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement

Evaluation

Evaluer l'adéquation des modèles aux objectifs métier :

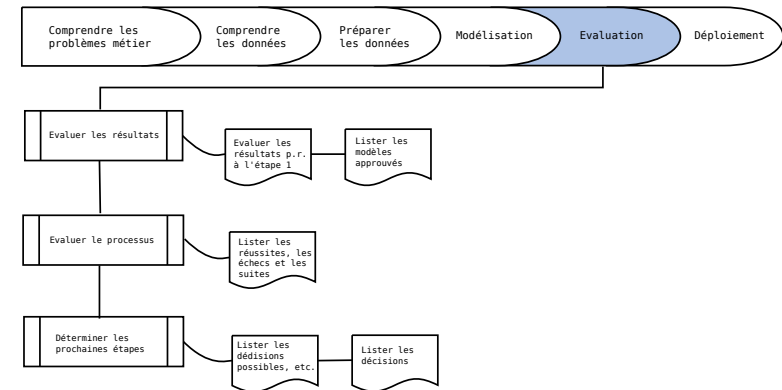
- évaluer les résultats
- évaluer le processus
- déterminer les prochaines étapes

↔ Produire des connaissances valides, intéressantes, etc. pour les objectifs métier.

5-1 évaluer les résultats

- évaluer les résultats par rapport aux critères métier et aux objectifs premiers
- identifier les raisons métier pour lesquelles les résultats sont satisfaisants ou non
- évaluer en situation réelle
- évaluer les résultats par rapport à des objectifs secondaires, des objectifs nouveaux révéler par le processus, etc.

↔ Evaluation par rapport aux objectifs initiaux mais aussi mise en évidence de challenges futurs.



5-1 évaluer les résultats

Délivrables : rapport d'évaluation, modèles

- lister les résultats par rapport aux critères métier et aux objectifs premiers
- mise en évidence de la réussite immédiate du projet, des lacunes à combler
- liste des modèles permettant d'atteindre les objectifs

↔ Liste des modèles approuvés.

5-2 évaluer le processus

- évaluer l'ensemble du processus
- vérifier qu'aucun facteur (attribut, tâche) important n'a été omis
- vérifier les critères de qualité
- vérifier l'utilité des données non utilisées
- vérifier la disponibilité future des données, des attributs, les risques associés

↔ Qualifier le processus, identifier des risques d'erreur.

5-3 déterminer les prochaines étapes

- le projet est-il terminé ?
- doit-on déployer de nouvelles itérations
- peut-on les déployer ?

↔ Peut-on déployer le projet ?

5-2 évaluer le processus

Délivrables : rapport sur le processus

- résumer le processus
- lister les tâches ayant été oubliées, ratées
- lister les tâches devant être à nouveau développées

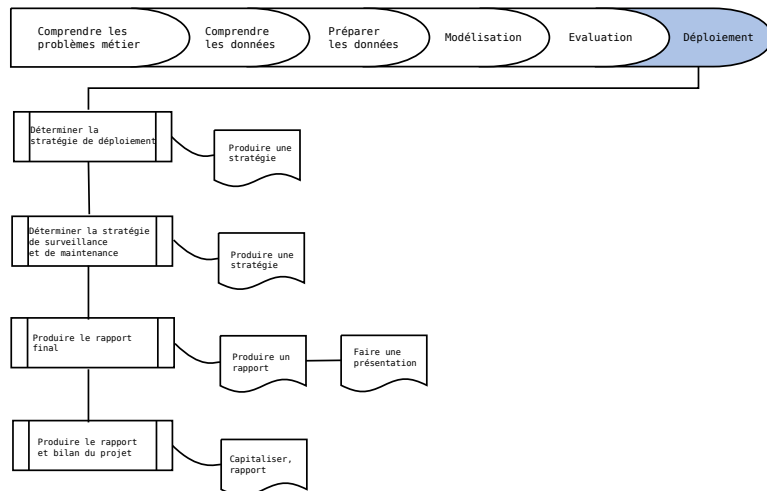
5-3 déterminer les prochaines étapes

Délivrables : rapport des actions possibles, décisions

- lister toutes les action possibles, les évaluer
- lister les décisions prises, la façon de les rendre opérationnelles, motivations

↔ Déploiement du projet et/ou nouvelles actions.

- 1 Méthodologies en fouille de données
- 2 CRISP-DM-1 : comprendre le(s) problème(s)
- 3 CRISP-DM-2 : comprendre les données
- 4 CRISP-DM-3 : préparer les données
- 5 CRISP-DM-4 : modélisation
- 6 CRISP-DM-5 : évaluation
- 7 CRISP-DM-6 : déploiement



Déploiement

Elaborer les actions nécessaires au déploiement des résultats, terminer le projet.

- une stratégie de déploiement
- une stratégie de surveillance et de maintenance
- un rapport final
- un bilan du projet

↔ Bref rentabiliser : intégrer les nouvelles connaissances aux processus métier afin d'améliorer l'activité et capitaliser.

6-1 stratégie de déploiement

Elaborer les actions nécessaires au déploiement des résultats.

- une stratégie de déploiement
- à partir de la phase d'évaluation
- généraliser si possible les procédures ayant été développées pendant le projet, i.e. capitaliser sur le processus

↔ Stratégie de mise en œuvre des résultats.

6-1 stratégie de déploiement

Délivrables : stratégie de déploiement

- résumé de la stratégie pour déployer les modèles dans l'organisation, dans les systèmes d'informations, dans les procédures, etc.
- détails avec toutes les étapes et sur la façon de les réaliser
- mesures pour évaluer les bénéfices

6-2 stratégie de surveillance et de maintenance

Délivrables : stratégie de surveillance

- résumé de la stratégie pour surveiller la bonne/mauvaise utilisation des modèles dans l'organisation, dans les systèmes d'informations, dans les procédures, etc.
- détails avec toutes les étapes et sur la façon de les réaliser
- mesures pour évaluer (évolution des objectifs métiers, évaluation des modèles dans le temps, évolution des données, etc.)

6-2 stratégie de surveillance et de maintenance

Elaborer les actions nécessaires pour surveiller le bon usage des modèles.

- une stratégie de surveillance et de maintenance
- identifier les critères à surveiller
- préparer la bonne utilisation des résultats

↪ Stratégie de mise en œuvre des résultats.

6-3 rapport final

Résumer de façon compréhensible le projet et les résultats.

- résumé du projet (si pas déployé)
- lister les expériences pertinentes du projet (si pas déployé)
- rapport final incluant les résultats si le projet est déployé

↪ Capitalisation, historique du projet, des résultats et valorisation.

6-3 rapport final

Délivrables : rapport, présentation

- rapport final, avec livrables, résumés, et résultats organisés
- présentation finale, avec livrables, résumés, et résultats organisés, discussion

↔ Convaincre également les clients, les collaborateurs, etc.

6-4 bilan du projet

Analyser le processus.


- lister ce qui s'est bien déroulé, mal déroulé
- identifier les causes

↔ Capitalisation pour de futurs projets similaires.

6-4 bilan du projet

Délivrables : rapport

- rapport sur les expériences importantes
- mettre en évidence ce qui est ré-utilisable pour des projets similaires
- satisfactions des utilisateurs, besoin en formation, etc.

-  Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2001). **Crisp-dm 1.0. step-by-step data mining guide.** Technical report, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands). <http://www.crisp-dm.org/CRISPWP-0800.pdf>.