

Data Mining

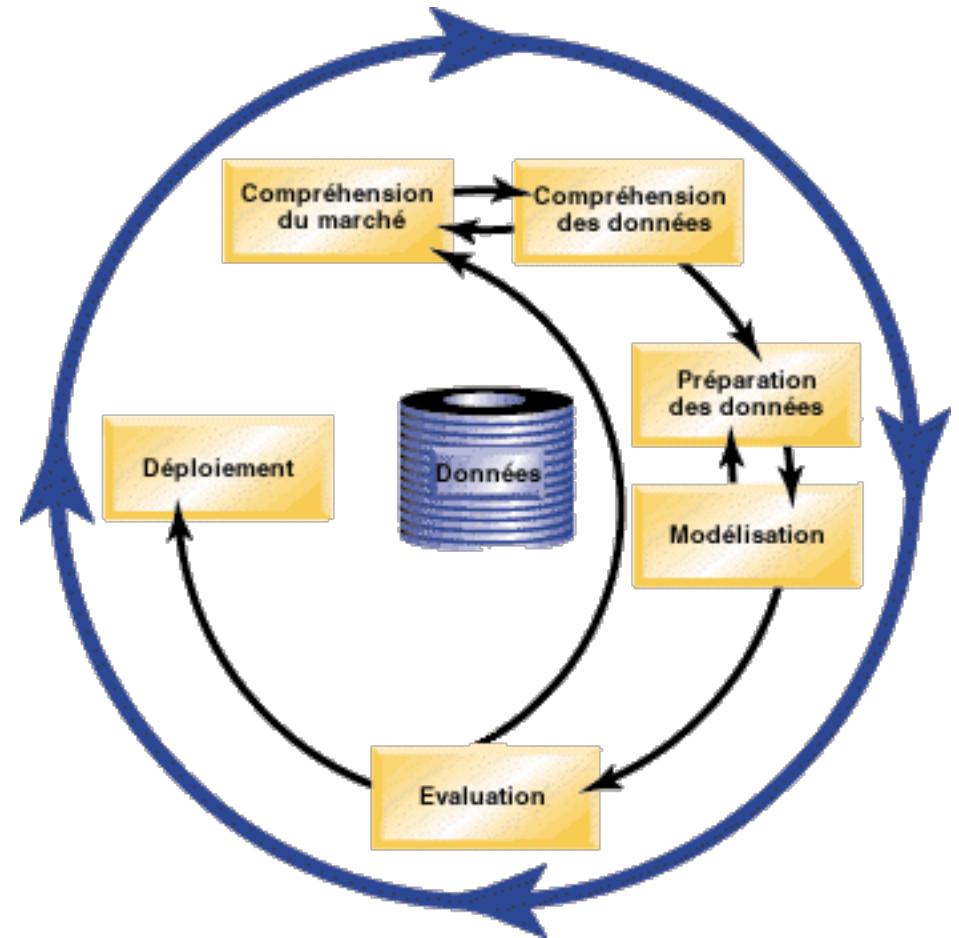
2. CRISP-DM

Nicolas Pasquier

<http://www.i3s.unice.fr/~pasquier>

Méthodologie CRISP-DM

- Modèle de cycle de vie à six phases
- Arcs : dépendances les plus importantes et fréquentes
- Séquence non stricte : dans la plupart des projets, on passe d'une phase à l'autre en fonction des besoins



CRISP-DM : Phase 1

- Adaptable : modèle aisément personnalisable
- Ces six phases couvrent la totalité du processus de Data Mining, y compris la façon de l'incorporer aux activités
- Compréhension du problème : étape primordiale qui vise à
 - déterminer des objectifs (commerciaux ou autres),
 - analyser la situation,
 - déterminer les objectifs en termes de Data Mining,
 - établir une planification du projet.

CRISP-DM : Phase 1

- Identifier les différents objectifs à atteindre
- Déterminer les facteurs importants qui peuvent influencer l'aboutissement du projet
- Inventorier :
 - les ressources,
 - les contraintes,
 - les hypothèses.
- Déterminer les objectifs de la fouille de données en termes techniques
- Description exhaustive de toutes les étapes à venir

CRISP-DM : Phase 2

- Compréhension des données : étape de sensibilisation à l'importance de bien maîtriser les ressources de données et leurs caractéristiques
- Elle aborde
 - la collecte de données initiale,
 - la description et l'exploration des données,
 - la vérification de la qualité de ces données.

CRISP-DM : Phase 2

- Charger les données
 - Rapport sur la nature, la localisation, les méthodes de récupération, les problèmes éventuels
- Examen rapide et superficiel des données
 - Les données satisfont-elles les conditions requises?
- Comprendre les données :
 - Utilisation de requêtes, outils de visualisation et de reporting.
 - Déterminer les attributs importants et leurs relations (redondantes)
 - Premiers résultats statistiques (graphiques, répartitions, etc.)
- Qualité des données
 - Données manquantes, erronées, incertaines?

CRISP-DM : Phase 3

- Préparation des données : transformation des données à explorer afin d'assurer leur adéquation à la problématique et la pertinence des connaissances extraites
- Phase de préparation comprend
 - la sélection,
 - le nettoyage,
 - La construction,
 - l'intégration,
 - Le formatage des données.

CRISP-DM : Phase 3

- Sélection des variables et instances
 - Selon l'objectif, la qualité des données, les contraintes techniques (catégorielles, numériques, etc.)
- Traitements des données bruitées
 - Correction des erreurs, estimation de valeurs
- Construction
 - Créer des valeurs dérivées, transformer des valeurs (normalisation, discrétisation)
- Combiner les données de diverses sources, supprimer les données redondantes
- Représentations adaptées, contraintes techniques

CRISP-DM : Phase 4

- Modélisation : élaboration des méthodes d'analyse qui seront utilisées pour extraire des connaissances à partir des données (cœur du processus)
- Cette étape consiste à
 - sélectionner des techniques de modélisation,
 - générer des conceptions de test,
 - construire et évaluer des modèles.

CRISP-DM : Phase 4

- Sélectionner au moins une technique d'extraction d'association, de classification et de clustering
- Définir les mécanismes pour évaluer la qualité et la validité du modèle
- Définition de jeux d'apprentissage et de test (échantillonnage)
- Exécution des algorithmes
 - Documenter les résultats (données, paramètres, modèles)
- Classer les modèles par intérêt
- Faire appel aux experts du domaine

CRISP-DM : Phase 5

- Évaluation : évaluer l'aide apportée par l'utilisation des modèles définis pour la concrétisation des objectifs poursuivis
- Cette étape aborde
 - l'évaluation des résultats,
 - la vérification du processus de Data mining,
 - la prise de décision des étapes à suivre.

CRISP-DM : Phase 5

- Évaluer l'adéquation des modèles aux objectifs métier
 - Rapport d'analyse des résultats selon les critères de succès
 - Approbation des modèles
- Vérification du processus
 - Vérifier qu'aucun facteur important n'a été oublié
 - Vérifier le respect des critères de qualité
 - Évaluer l'utilité potentielle des données non utilisées
 - Les données utilisées seront elles toujours disponibles?
- Décider des étapes suivantes
 - Le projet est il prêt à être déployer?
 - Nouvelles itérations du processus nécessaires?

CRISP-DM : Phase 6

- Déploiement : étape de rentabilisation des efforts déployés
- Objectif : intégrer les nouvelles connaissances aux processus quotidiens pour résoudre le problème initial / améliorer l'activité
- Elle comprend
 - le déploiement du plan,
 - la surveillance et la maintenance,
 - la production d'un rapport final,
 - la révision du projet.

CRISP-DM : Phase 6

- Analyser les évaluations
 - Comment déployer les modèles dans l'organisation?
 - Comment analyser les bénéfices du modèle?
- Rapport final
 - Résumer le projet et l'expérience acquise
 - Présenter de façon compréhensible les résultats du data mining
- Bilan du processus
 - Analyser le processus et observer ce qui c'est bien ou mal déroulé
 - Les utilisateurs sont-ils pleinement satisfaits ? Ont-ils besoin de support ?