Analyse comportementale basée sur de l'apprentissage et de la vérification.

Directeur de thèse : Philippe Leray – équipe DuKe

Encadrant : Benoît Delahaye – équipe Aelos

Dimitri ANTAKLY

LE Groupe GFI Informatique

14 000
COLLABORATEURS

16 PAYS

FRANCE
ESPAGNE
PORTUGAL
BELGIQUE
SUISSE
LUXEMBOURG

ANGLETERRE
POLOGNE
ROUMANIE
MAROC
CÔTE D'IVOIRE
ANGOLA

USA
MEXIQUE
COLOMBIE
BRÉSIL

5 VALEURS

AMBITION
INNOVATION
ENGAGEMENT
ESPRIT D'ÉQUIPE
RESPONSABILITÉ SOCIÉTALE

1015
MILLIONS D'EUROS
DE CHIFFRE D'AFFAIRES
2016

16 CENTRES DE SERVICES PARTAGÉS

EN FRANCE
- LILLE
- LYON
- NANTES
- TOULOUSE
- MEUDON

À L'INTERNATIONAL
- ALICANTE (Espagne)
- LISBONNE (Portugal)
- COVILHA (Portugal)
- CASABLANCA (Maroc)
- VARSOVIE (Pologne)
- POZNAŃ (Pologne)
- LUBLIN (Pologne)
- PUNE (Inde)
- SÃO PAULO (Brésil)
- BOGOTA (Colombie)
- MACAO (APAC)

6 MÉTIERS
- CONSULTING
- APPLICATION SERVICES
- INFRASTRUCTURE SERVICES
- BUSINESS SOLUTIONS
- SOFTWARE
- SAP

Gfi Informatique est partenaire majeur du Paris Saint-Germain Handball

\+ d'innovation de proximité d'industrialisation

5 SOLUTIONS MÉTIERS
- ASSURANCE
- DISTRIBUTION-SERVICES
- SANTÉ-SOCIAL
- SECTEUR PUBLIC
- TÉLÉCOM

6 SECTEURS D'ACTIVITÉS
- BANQUE-FINANCE-ASSURANCE
- INDUSTRIE-AÉROSPATIAL-TRANSPORT
- SECTEUR PUBLIC
- TÉLÉCOM-MEDIA-ENTERTAINMENT
- ÉNERGIE-UTILITIES-CHIMIE
- DISTRIBUTION-SERVICES

8 PRACTICES GROUPE
- Cybersécurité
- DevOps
- Digital Banking
- IoT
- Omni Commerce
- Smart Cities
- Transformation Digitale
- Usine 4.0

ET LA BOURSE ?

L'action Gfi Informatique est cotée à la bourse de Paris, Euronext.
Code ISIN : FR 0004038099

www.gfi.world

gfi NEW CHALLENGES, NEW IDEAS

Centre d'Innovation et d'Expertise : un centre de compétences qui s'appuie sur un savoir-faire

3 Orientations
- EXPERTISE
- CONSEIL Technologique
- INNOVATION

3 Sites
- NANTES
- PARIS
- PACA

3 Vocations
- INNOVATION
- VITRINE TECHNOLOGIQUE
- EXPERTISE

50
Key People
Référents
Expert IT

Microsoft Partner
Gold Independent Software Vendor (ISV)
Silver Desktop
Silver Server Platform
Silver Virtualization
Silver Systems Management
Silver Identity and Security

ORACLE

Symantec. Registered Partner

CISCO

SAP Partner

vmware

CITRIX

LoginPeople

EMC² VELOCITY² AFFILIATE PARTNER

NUTANIX

Watchful

gfi NEW CHALLENGES, NEW IDEAS

Cloud Experience

open source

L'ambition du CIE :

« Être l'expert dans toute transformation de vos infrastructures à travers des solutions innovantes»

PROGRAMME IP25 2016

3 Practices Technologiques
- *Datacenter*
- *Sécurité*
- *End-User Computing*

5 Piliers
- Expertise et consulting
- Architecture d'infrastructure
- Design de Transformation
- Optimisation et support de Production
- Solutions innovantes

gfi People Map
gfi MaXiM
gfi SPIM
gfi Keenaï
gfi Security Service

6 Domaines
- → Réseau et Sécurité
- → Cloud
- → Communication et Digital
- → Gestion des outils industriels
- → Environnement Data Center
- → Poste de Travail

Contexte

- Les solutions actuelles essentiellement périmétriques, ne répondent plus aux besoins de cybersécurité
- En chiffres : Cybercriminalité représente 41% de la criminalité devant la drogue (29%) et le trafic d'arme (21%)
- L'analyse statistique ne permet pas de répondre à ces nouvelles menaces

Objectif de ce travail

- Contribuer à l'analyse actuelle par des moyens de ML
- Etablir un lien entre les techniques d'apprentissage (RBD) et le SMC
- Se joindre aux uses cases GFI
    - Des approches novatrices dans le SIEM
    - Moteur d'analyse comportementale

1. **Réseaux Bayésiens et réseaux Bayésiens dynamiques**

2. **Data/Process Mining**

3. **Analyse syntaxique (Inférence grammaticale)**

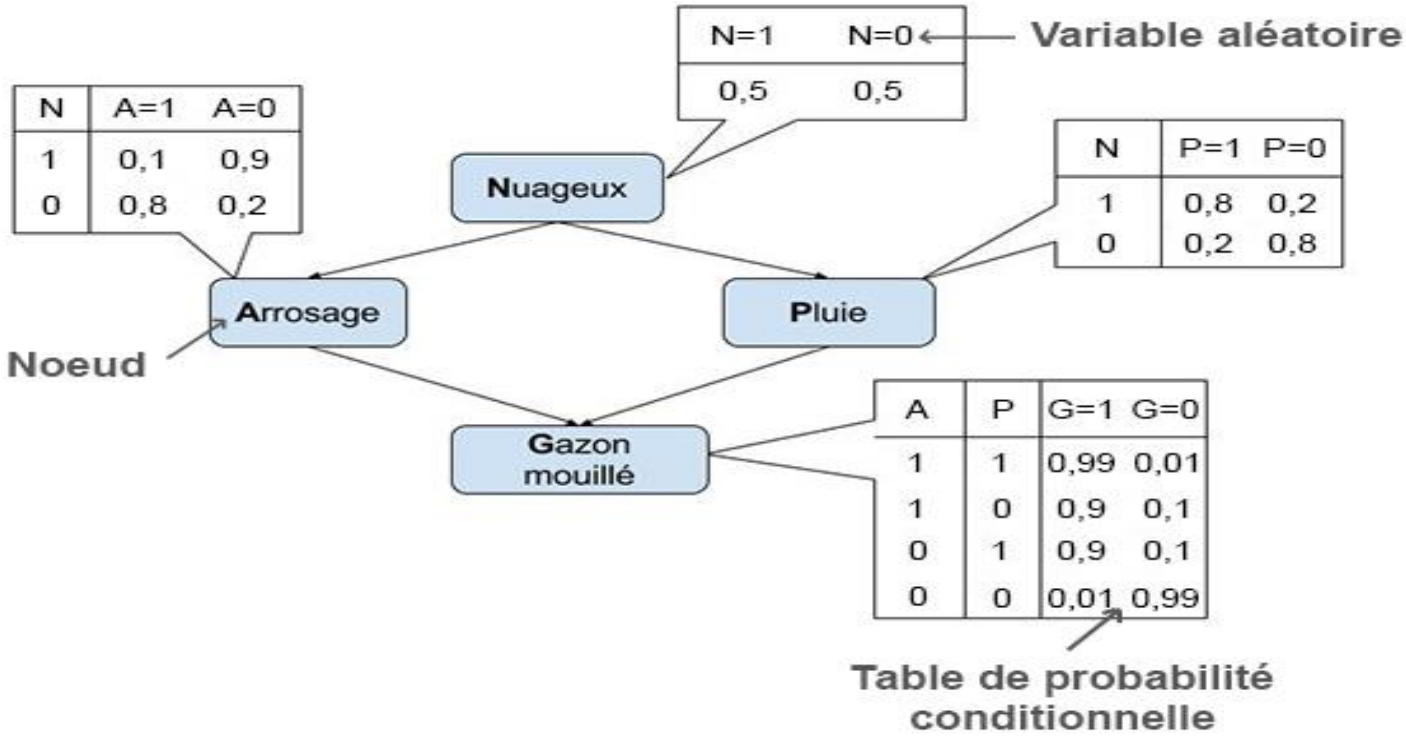
4. **Statistical Model Checking (SMC)**

II – Etat de l'art – RB et RBD 1/2
RB

Un réseau Bayésien $R(G, P)$ est défini par :

- G, un DAG (Directed Acyclic Graph)
- P, l'ensemble des probabilités de chaque nœud conditionnellement à l'état de ses parents.

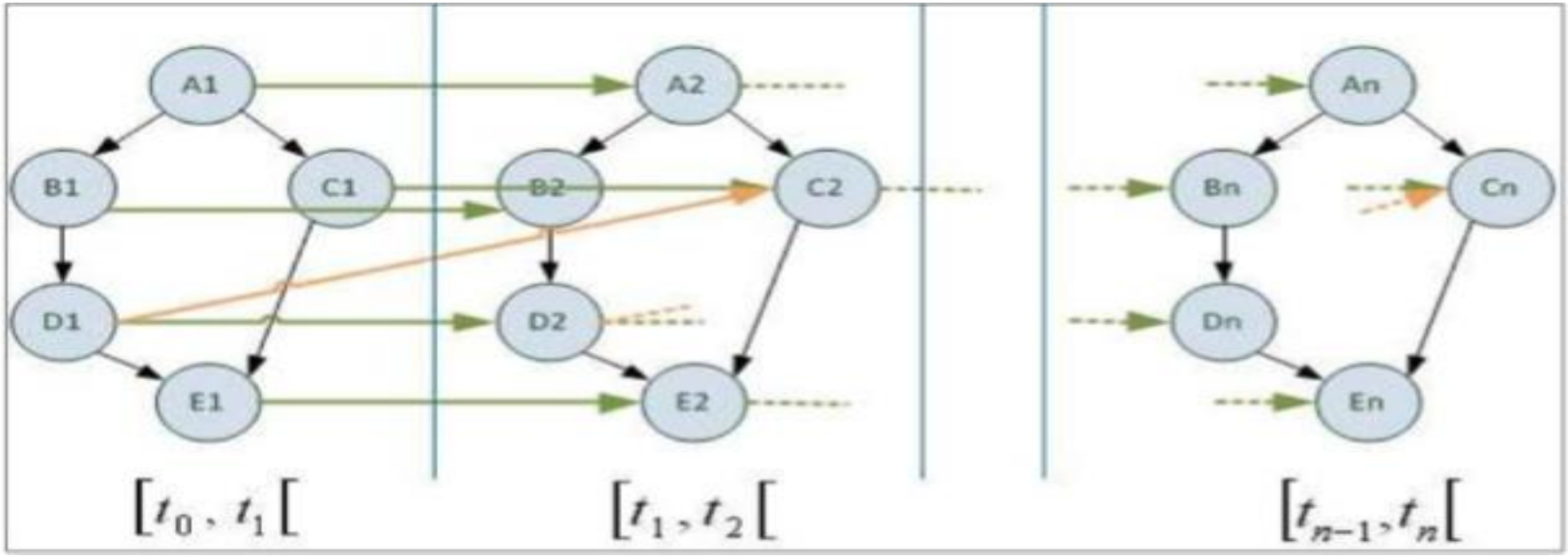
Notons X_i l'ensemble des nœuds, alors : $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | P_a(X_i))$.



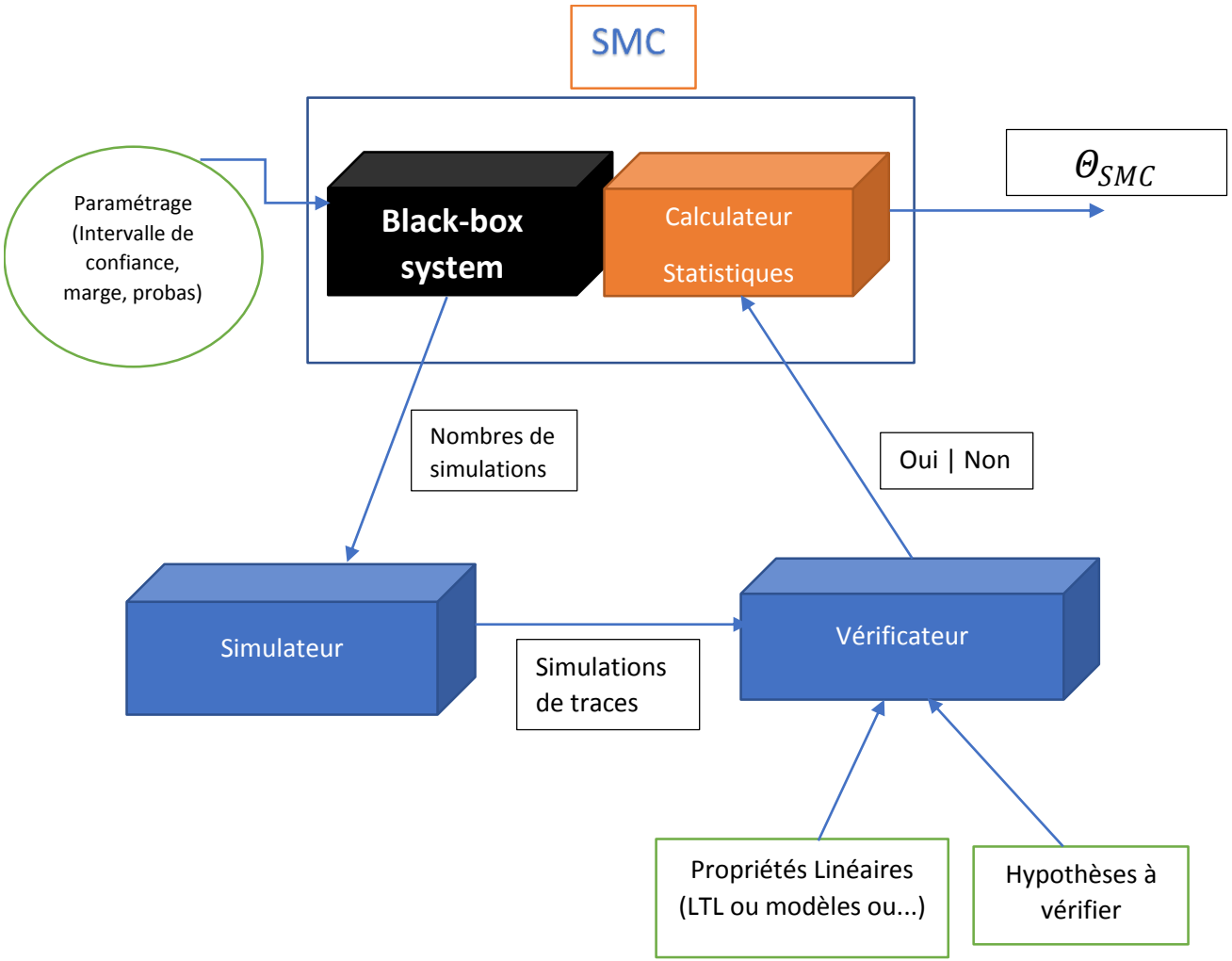
II – Etat de l'art – RB et RBD 2/2
RBD

Les questions traditionnelles, sont réparties sur plusieurs catégories :

- Filtrage : Calculer $P(X_t \mid Y_{0:t})$ AVEC $X_t = \{X_{it}\} = l'ensemble\ de\ variables\ aléatoires\ au\ temps$ « t ».
- Lissage : Calculer $P(X_{t-l} \mid Y_{0:t})$
- Prédiction : Calculer $P(X_{t+h} \mid Y_{0:t})$
- Décodage : Calculer $arg\ max\ P(X_{0:t} \mid Y_{0:t})$
- Classification : Calculer $P(Y_{0:t})$



II – Etat de l'art – SMC 1/2



II – Etat de l'art – SMC 2/2

Avantages :

- Les algorithmes requièrent uniquement que le système soit exécutable afin de le simuler (pas besoin de modèle prédéfini)
- Facilement parallélisables, donc s'applique sur du Big Data

Inconvénients :

- Il faut que le système ne contienne pas du non-déterminisme
- Propriétés bornés dans le temps
- Inefficace sur de très faibles probabilités
- Des échantillons plus larges, si plus de précision est requise

Les applications du Data Mining :

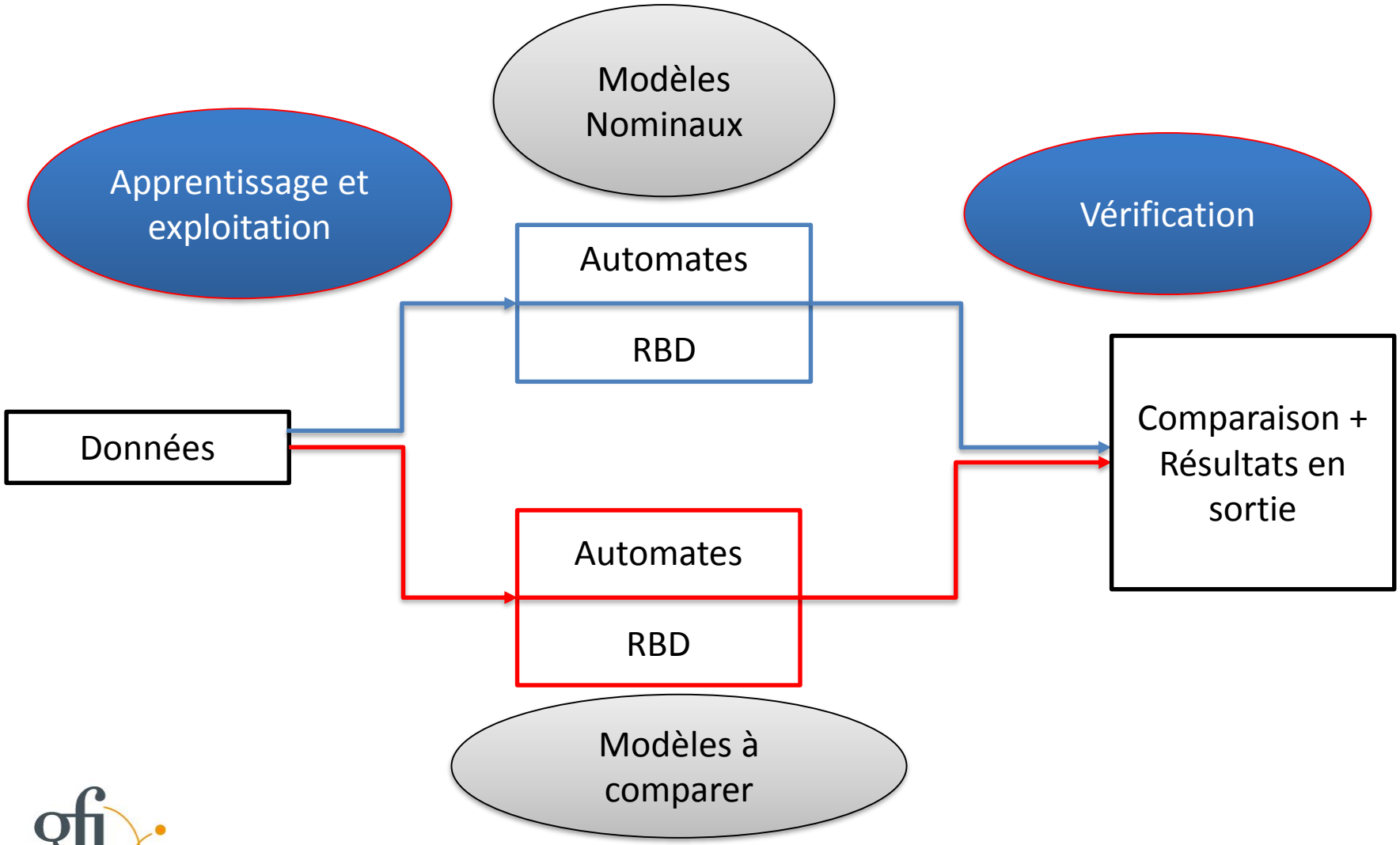
- Analyse de données et aide à la décision,
- Analyse de marché,
- Marketing ciblé, gestion des relations client, analyse des achats des clients, ventes croisées, segmentation du marché ;
- Analyse de risque,
- Détection de fraudes (Utiliser les données historiques pour construire des modèles de comportements frauduleux puis utiliser les techniques de datamining pour retrouver des instances similaires).

L'objectif principal du Process Mining est de répondre à ces questions :

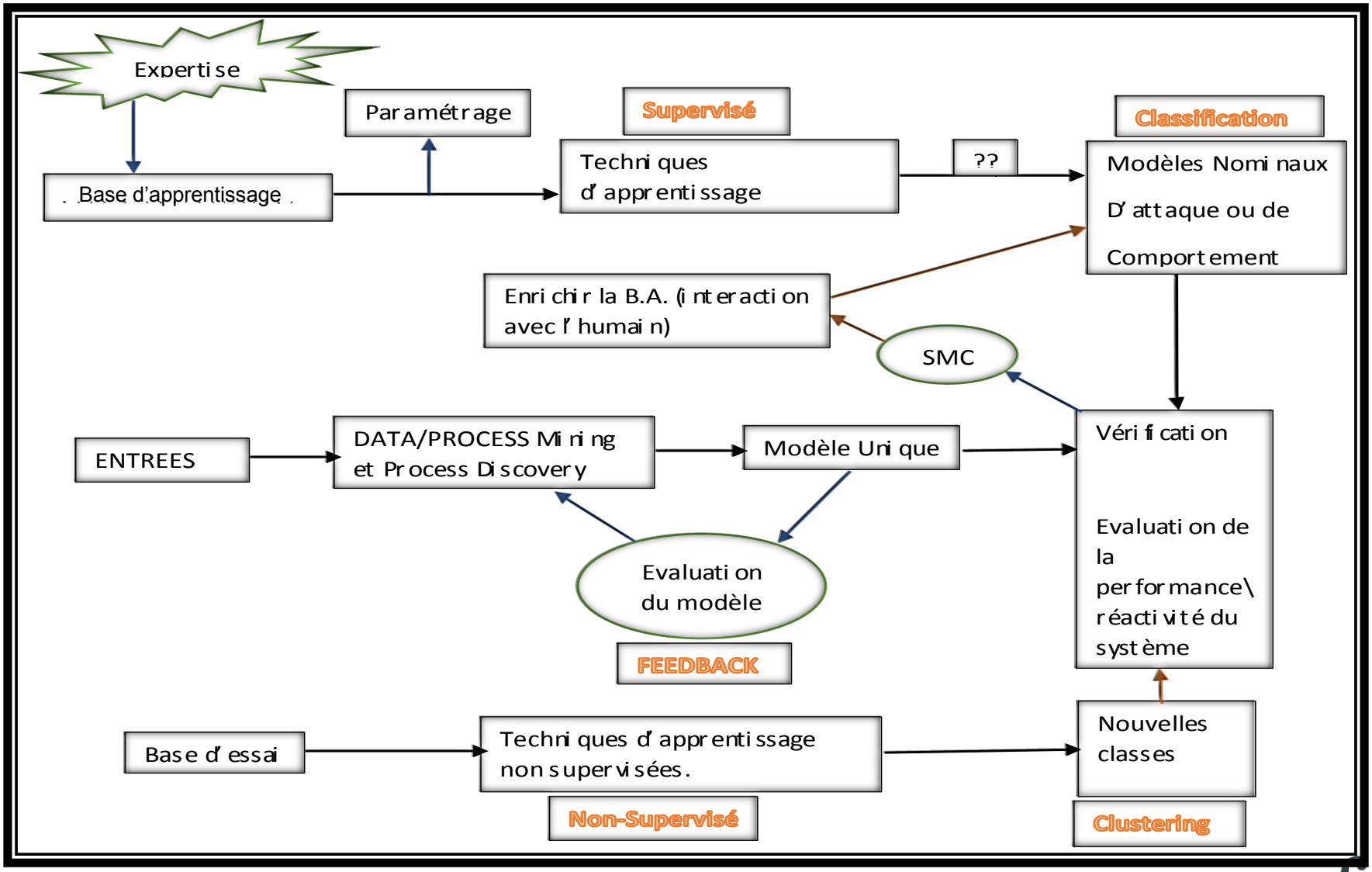
- Qu'est ce qui s'est vraiment produit dans le passé?
- Pourquoi ça s'est passé ?
- Qu'est ce qui va probablement se produire dans le futur ?
- Quand et pourquoi les organisations et les personnes ont un comportement déviant ?
- Comment mieux contrôler le procédé ?
- Comment reconstruire un procédé pour améliorer sa performance ?

- **Informed learners :**
  - PTA (ou Prex Tree Acceptator)
  - GOLD's Algorithme
- **Techniques d'IA :**
  - GA
  - Principe MDL
  - Heuristiques
- **Apprentissage de PFA :**
  - ALERGIA (et ses extensions)
  - Les heuristiques (MDI)

III – Le contexte générique du moteur à construire



Première approche détaillée ..\Plan de Travail.docx



III – Perspective de GFI

Les fonctionnalités attendues

- Un apprentissage intelligent basé sur le principe de « Desire Lines »
- Faire des préventions adaptatives, en d'autres termes, un comportement inconnu ou nouveau ou bizarre, n'est pas nécessairement malveillant
- Les préventions sont probabilistes et non pas déterministes, donc plus réaliste

Les bénéfices attendus

- Limiter, au bon moment, les comportements « bizarres », de sorte de ne pas les interdire complètement mais les restreindre (il s'agit de la Précision, en termes de qualité)
- Améliorer l'anticipation sur les cas à problèmes
- User-friendly
- Augmentation de la QoS