# Système adaptatif de détection d'intrusions dans des serveurs Web

**Équipe DREAM - IRISA/INRIA**
Guyet Thomas
René Quiniou
Wei Wang
Marie-Odile Cordier

# Systèmes de détection d'intrusions

- Objectif général :
  - Protéger les informations contenues sur un serveur Web
  - Contrecarrer les attaques visant l'inactivation du serveur
- But de l'IDS : Détecter et identifier des utilisations 'malveillantes' d'un serveur web
  - Au plus tôt
  - De manière robuste et précise
  - **Des attaques connues ou non**

# IDS à partir de log Apache

- À partir des logs d'accès à un serveur Apache
  - Flot de données structurées (~ 10 requêtes par secondes)


crawl-66-249-66-136.googlebot.com - - [09/May/2007:21:42:39 +0200] "GET /sloop/David.Coudert/Biblio/Year/index.php3?url=Biblio/Year/2005.complete.html HTTP/1.1" 200 176 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
pluviose.inrialpes.fr - - [09/May/2007:21:42:27 +0200] "GET /axis/cbrtools/manual/first_page.html HTTP/1.1" 304 - "-" "NG/2.0"
eruessel.stusta.mhn.de - - [09/May/2007:21:42:27 +0200] "GET /lemme/Hanane.Naciri/these/mathml/images/box.gif HTTP/1.0" 200 183 "http://www-sop.inria.fr/lemme/Hanane.Naciri/these/mathml/main.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"
eruessel.stusta.mhn.de - - [09/May/2007:21:42:27 +0200] "GET /lemme/Hanane.Naciri/these/mathml/images/archmathml.gif HTTP/1.0" 200 6665 "http://www-sop.inria.fr/lemme/Hanane.Naciri/these/mathml/main.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"
crawl-66-249-66-136.googlebot.com - - [09/May/2007:21:42:29 +0200] "GET /reves/Xavier.Granier/GIS/html/class_clusterizer.html HTTP/1.1" 200 7602 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
pluviose.inrialpes.fr - - [09/May/2007:21:42:37 +0200] "GET /cgi-bin/fom.cgi?_insert=answer&cmd=addItem&file=1&keywords=%3f HTTP/1.1" 302 17 "-" "NG/2.0"
pluviose.inrialpes.fr - - [09/May/2007:21:42:37 +0200] "GET /cgi-bin/fom.cgi?_insert=answer&cmd=addItem&file=1&keywords=%3f HTTP/1.1" 302 17 "-" "NG/2.0"
ferrier.biac.duke.edu - - [09/May/2007:21:43:56 +0200] "GET /asclepios/personnel/Pierre.Fillard/softwares/FiberTracking/DTITrack2005_manual.pdf HTTP/1.1" 206 585350 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 2.0.50727; InfoPath.1)"
crawl-66-249-66-136.googlebot.com - - [09/May/2007:21:42:27 +0200] "GET /acacia/ESSI/Images/%3FS=A&h=437&w=822&sz=9&hl=en&start=7/Fig-cycle-conception-IHM-LN.fm/Bad-printer-icon.fm/proprietes-IHM-LN.ps/osf.TIFF.gz/Logo-Moduel-IHM-design.ppt/arbre-des-colecticiels/Bad-printer-icon.fm/Bad-printer-icon.fm/Bad-printer-icon.ps/Fig-Modele-activite-Norman.fm/?D=A HTTP/1.1" 200 11880 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
eruessel.stusta.mhn.de - - [09/May/2007:21:42:27 +0200] "GET /lemme/Hanane.Naciri/these/mathml/images/figue.gif HTTP/1.0" 200 1540 "http://www-sop.inria.fr/lemme/Hanane.Naciri/these/mathml/main.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"

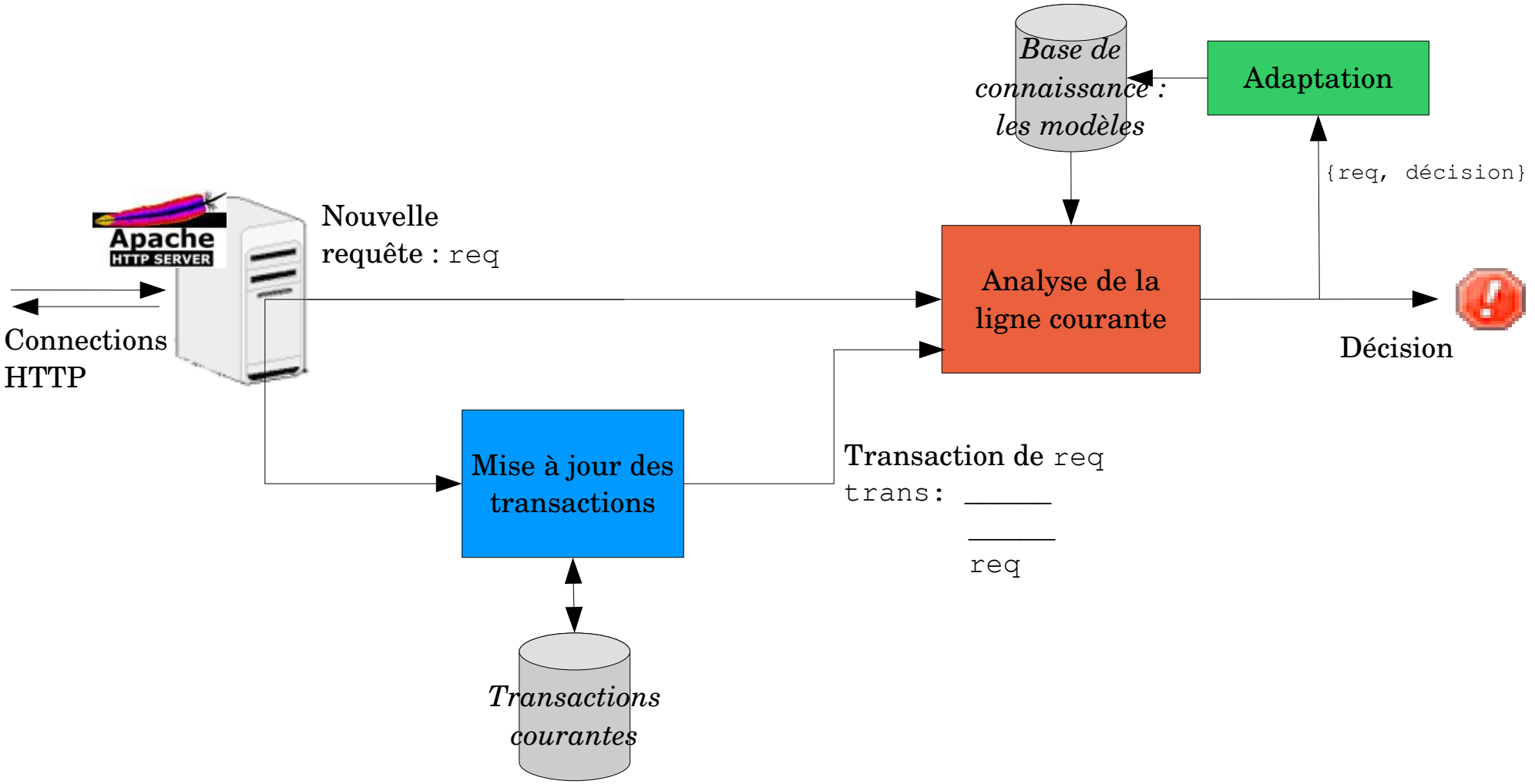
# IDS à partir de log Apache

- À partir des logs d'accès à un serveur Apache
  - Flot de données structurées (~ 10 requêtes par secondes)
  - Deux types de structure:
    - Les requêtes
    - Les transactions: Une transaction contient les requêtes faites par un même client
- Utilisation de modèles
  - D'intrusion
  - De comportement normal

# IDS à partir de log Apache

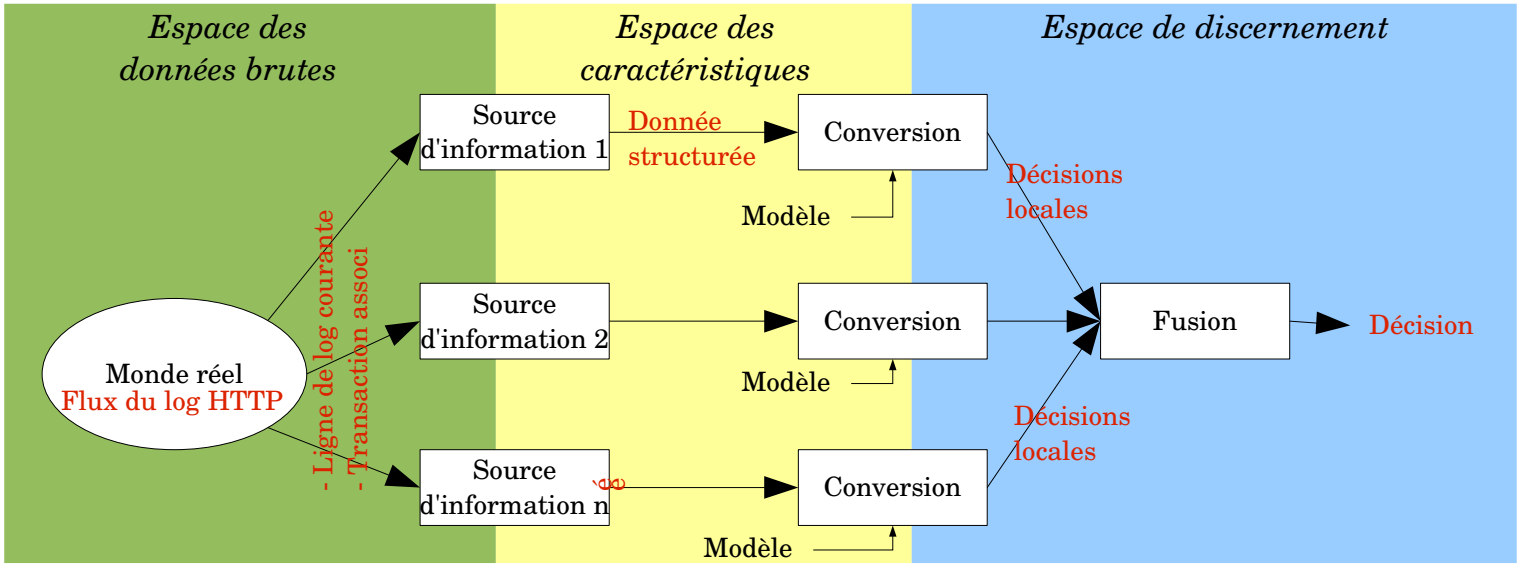
- Objectif : Adapter dynamiquement les modèles pour:
  - Améliorer les performances du système sur les attaques connues,
  - Découvrir de nouvelles attaques.

# Architecture générale de la détection d'intrusions



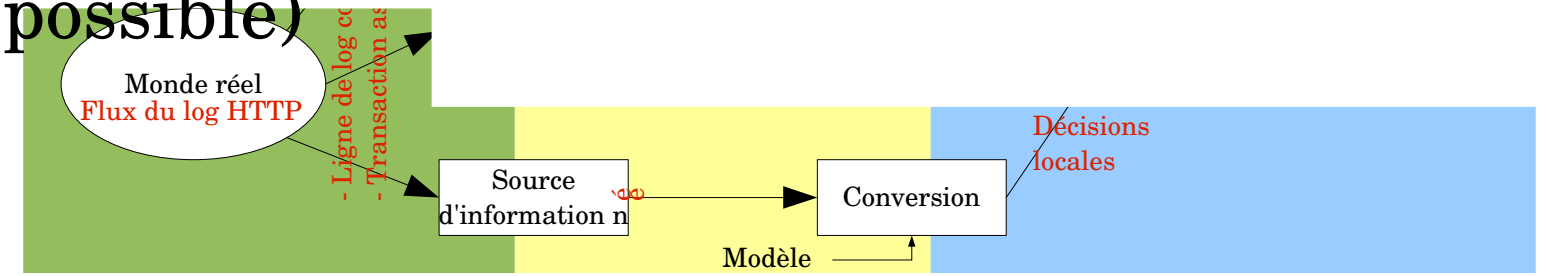
# Analyse de la ligne courante

- Fusion d'informations : utilisation de plusieurs indices pour identifier une intrusion
  - *Distribution de caractères de l'URL*
  - *Le status code (404 : erreur, 200 : ok)*
  - *Distribution de tokens de l'URL*
  - …



# Analyse de la ligne courante

- Conversion des informations
  - Étape qui transforme les données d'une source d'information dans l'espace de discernement (Ω)
    - Utilise un "modèle" (*ex: modèles de distribution de caractères pour une requête intrusive*)
  - Ω : ensemble des réponses possibles (non-nécessairement disjointes)
  - Une décision = un ensemble de masses (une par réponse possible)



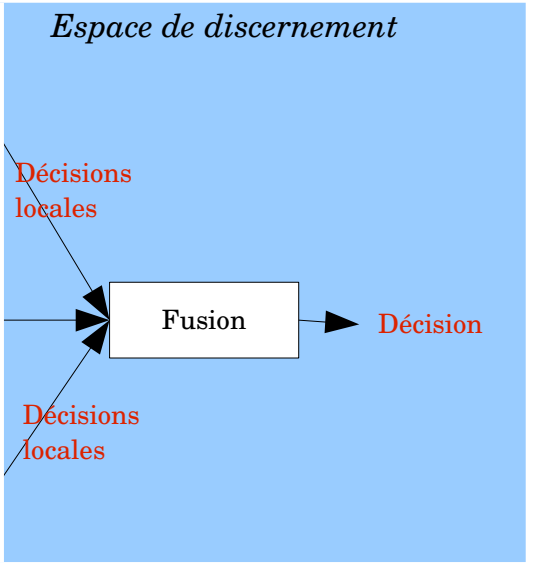
# Analyse de la ligne courante

- Conversion des informations
- $\Omega = \{intrusion, normalité, inconnue\}$
- Décision $d$:
  - Masse d'intrusion : $m(I)$
  - Masse de normalité : $m(N)$
  - Masse d'inconnue : $m(U)$ $\Rightarrow m(I) + m(N) + m(U) = 1$
  - Flag d'indécidabilité

# Analyse de la ligne courante

- Fusion de décisions : construit une décision globale à partir des décisions locales
  - $d = d_A + d_B$
  - Méthode de Shafer, moyenne des poids, ANN, ...
- Shafer
  - Fusion de deux sources *1* et *2* :
  - La masse de toute décision *A* est calculée par la formule :

$$m(A) = (m_1 \oplus m_2)(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}.$$



# Analyse de la ligne courante

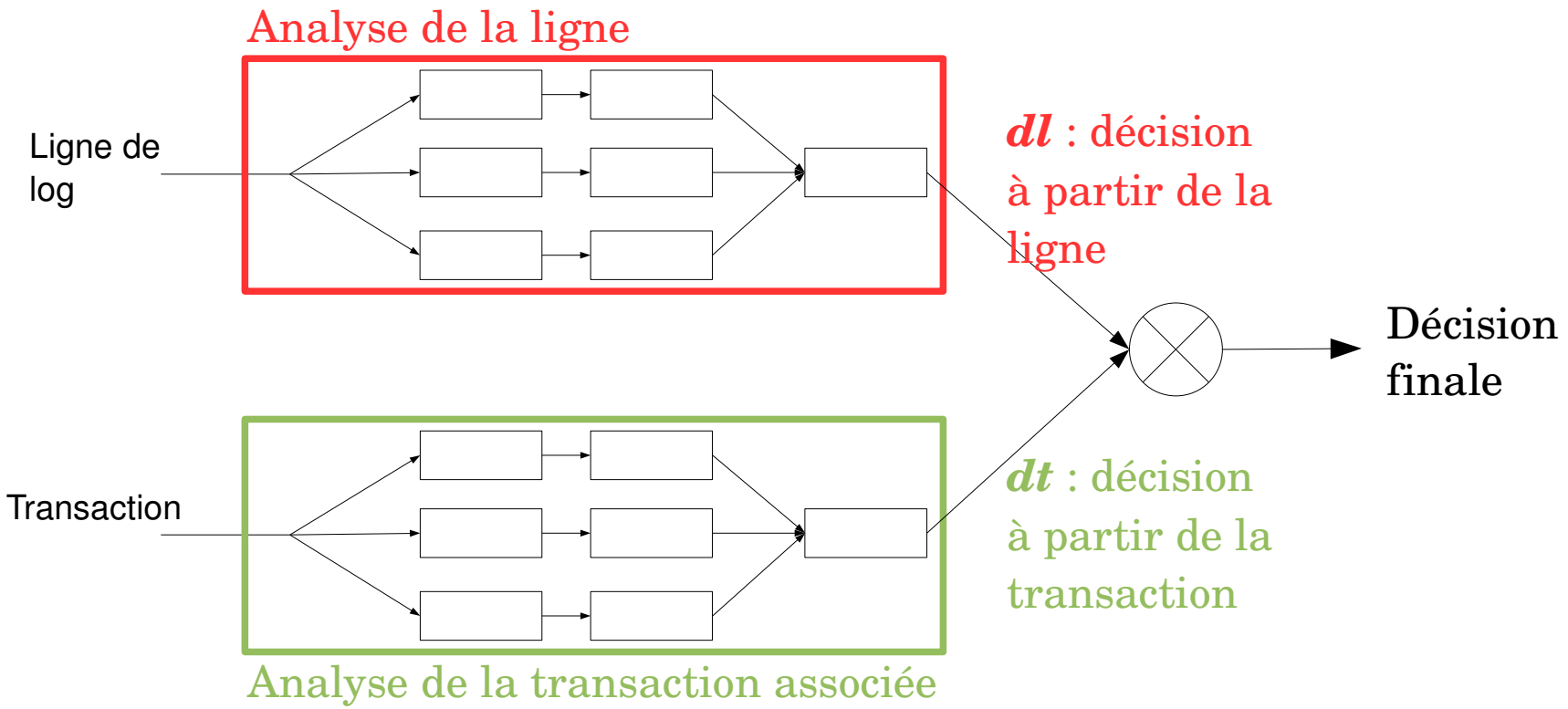
- Fusion dans notre espace de discernement :
  - On ne fusionne que les décisions provenant de sources *décidables*
  - Ω est une partition, d'où une simplification :

$$\forall A \in \Omega,\ m(A) = \frac{m_1(A)m_2(A)}{1 - \sum_{B \cap C = \emptyset} m_1(B)m_2(C)}$$

$$\sum_{B \cap C = \emptyset} m_1(B)m_2(C) = m_1(N)m_2(I) + m_1(N)m_2(U) + m_1(U)m_2(I) + m_1(U)m_2(N) + m_1(I)m_2(N) + m_1(I)m_2(U)$$

# Analyse de la ligne courante

- Second niveau de fusion

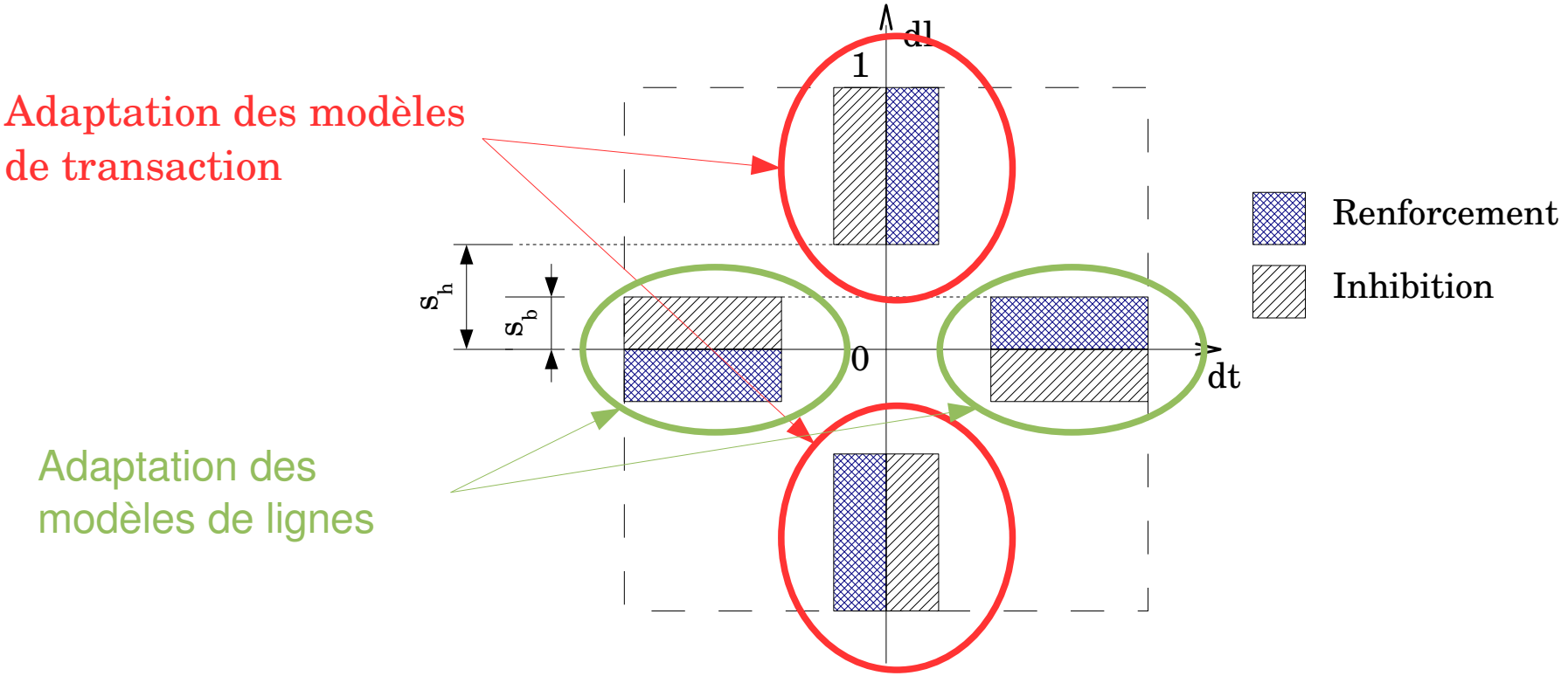


# Choix d'un besoin d'adaptation

- À partir de quels requêtes doit-on adapter les modèles ?
  - Avec un flux de données : on dispose de peu d'informations à confronter !
- Hypothèse : les requêtes d'une transaction doivent toutes être en accord avec la décision sur la transaction
- Heuristique : On adapte les modèles de lignes de log lorsque la décision *dl* est faible ($< s_b$) et que la décision *dt* est forte ($> s_h$), et réciproquement.

# Choix d'un besoin d'adaptation

- Illustration de l'heuristique

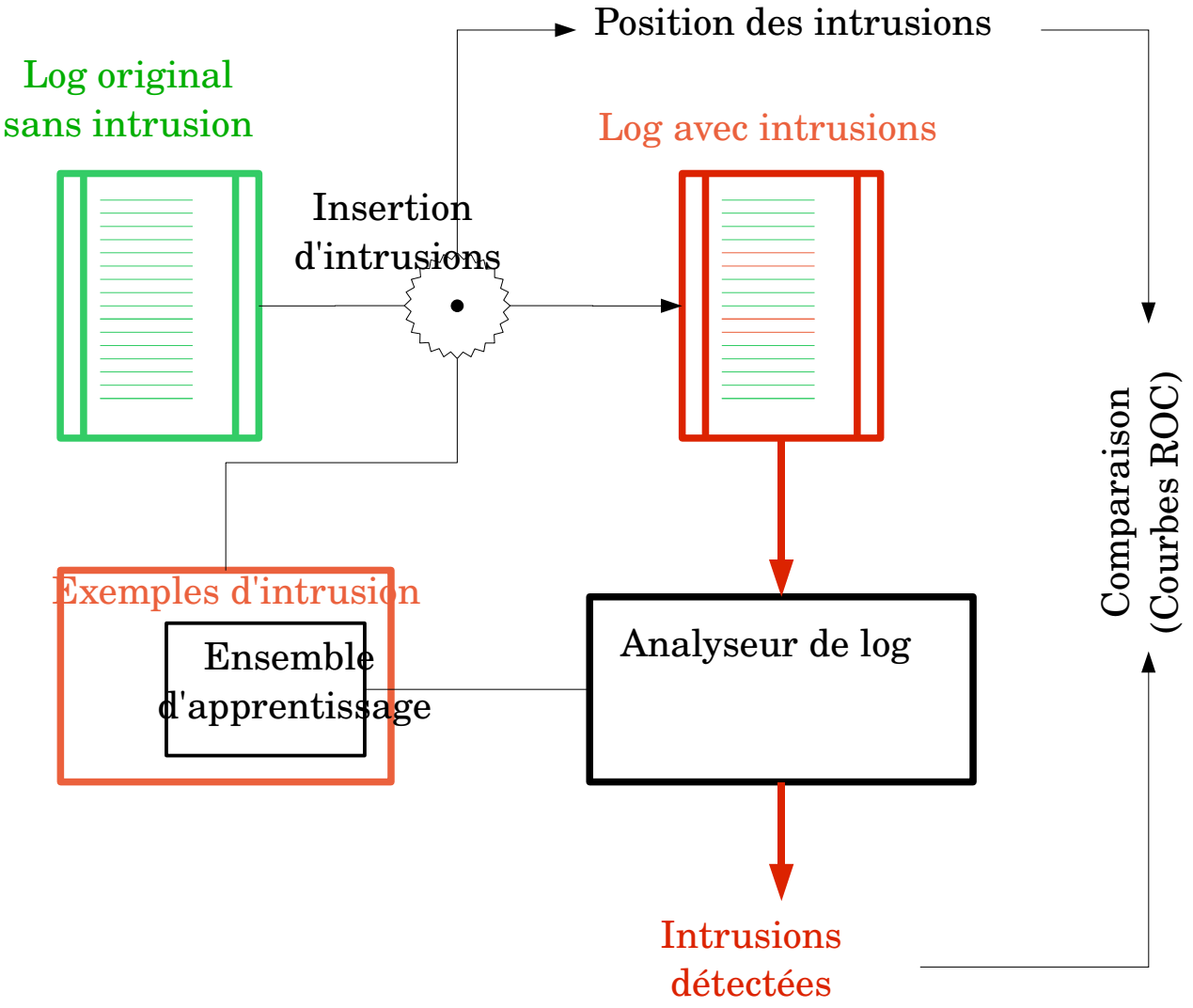


# Adaptation

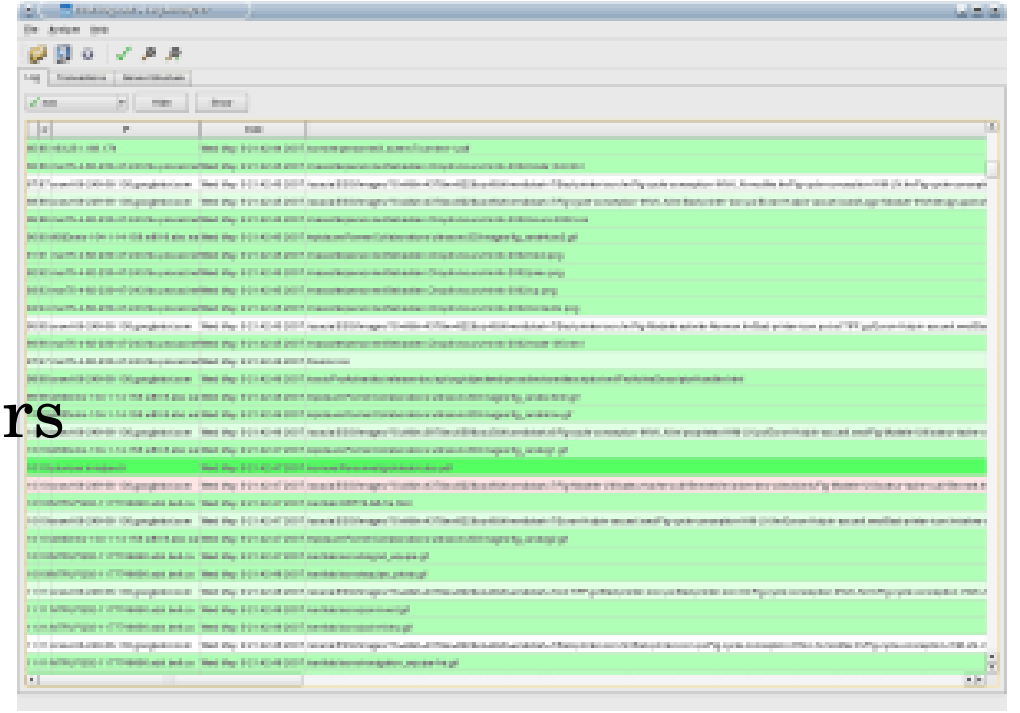
- Le choix d'une adaptation indique les modèles à adapter
  - Modèles de transaction ou de ligne
  - Modèles d'intrusion et/ou de normalité

- Chaque modèle dispose de sa propre fonction d'adaptation
  - Adapte le modèle à partir de la ligne courante
  - *Ex: Adaptation de la distribution de caractères de l'URL*

# Expérimentations



# Implémentation



- LogAnalyzer (C++)
  - Outil d'analyse interactif
  - Permet de traiter les fichiers à la volée

- Fonctionnalités
  - Visualisation, filtrage, statistiques, ...

**http://www.irisa.fr/dream/LogAnalyzer/**

# Implémentation

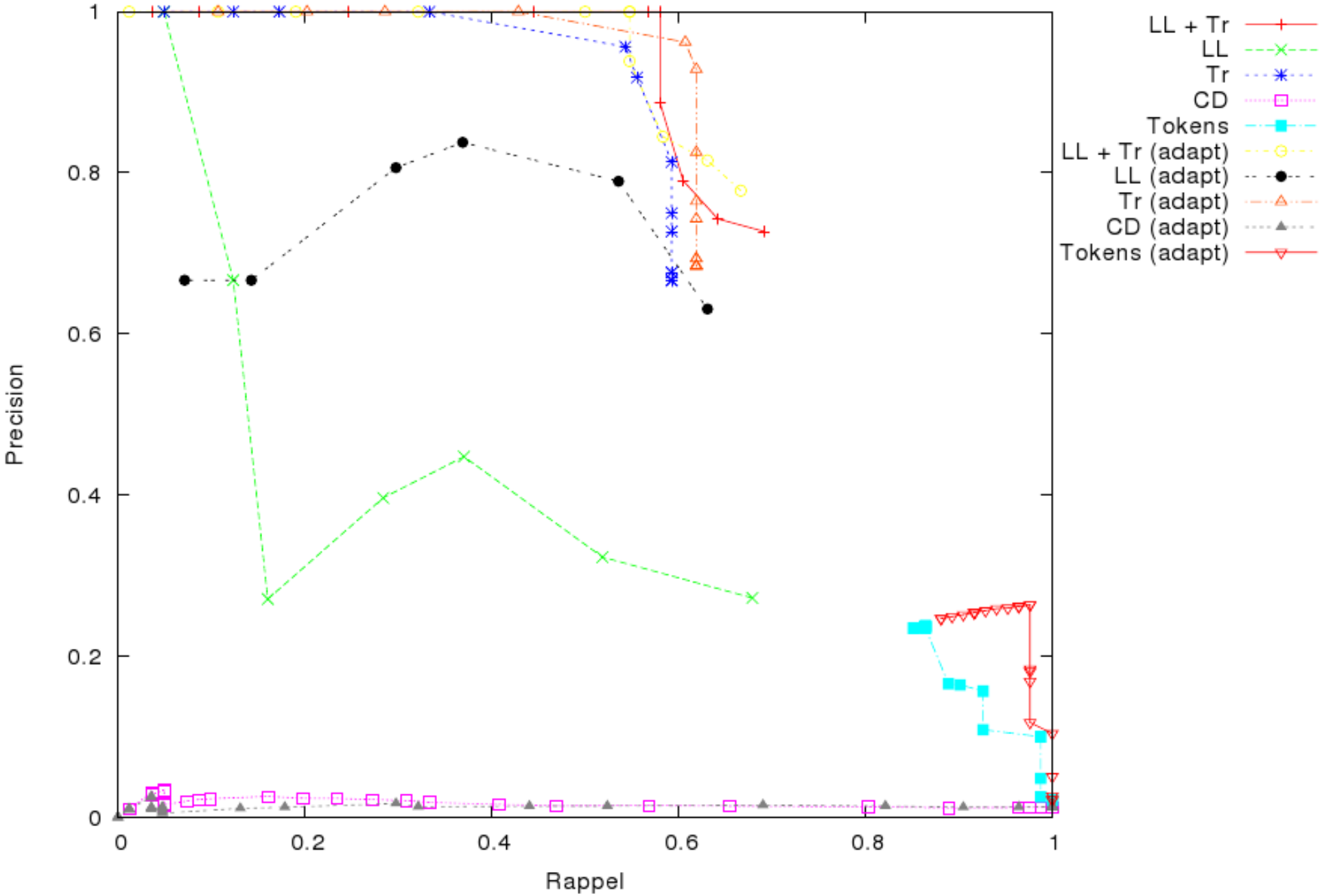
- Différents indices implémentés
- 8 indices sélectionnées pour mettre en évidence l'adaptation
  - Distribution de caractères (lignes et trans., intrusions et normales)
  - Distribution de tokens (lignes, intrusions et normales)
  - Proportion de code erreur (trans., intrusions et normales)

# Résultats

| | Shafer | Moyenne |
|---|---|---|
| Nb d'adaptation | 121.04 ($\pm$ 97.94) | 139.19 ($\pm$44.36) |
| Adaptations correctes (%) | **0.95** ($\pm$0.02) | 0.93 ($\pm$0.03) |
| Expérimentations détectant la nouvelle intrusion (%) | **0.35** | 0.27 |

| | Shafer | | Moyenne | |
|---|---|---|---|---|
| | Avec adaptation | Sans adaptation | Avec adaptation | Sans adaptation |
| Temps (s.) | 12.68 ($\pm$ 0.46) | **11.83** ($\pm$ 0.38) | 10.12 ($\pm$ 0.22) | **9.58** ($\pm$ 0.19) |
| Rappel | **0.94** ($\pm$ 0.03) | 0.93 ($\pm$ 0.04) | **0.50** ($\pm$ 0.26) | 0.49 ($\pm$ 0.33) |
| Precision | **0.81** ($\pm$ 0,25) | 0.78 ($\pm$ 0.31) | **0.93** ($\pm$ 0,09) | 0.91 ($\pm$ 0.09) |
| F-mesure | **0.85** ($\pm$ 0.18) | 0.81 ($\pm$ 0.22) | **0.61** ($\pm$ 0.28) | 0.55 ($\pm$ 0.35) |

# Résultats


Precision
Rappel
0
0.2
0.4
0.6
0.8
1
LL + Tr
LL
Tr
CD
Tokens
LL + Tr (adapt)
LL (adapt)
Tr (adapt)
CD (adapt)
Tokens (adapt)

# Perspectives

- Développer d'autres modèles pour rendre les décisions plus robustes
- Évaluer le système sur des intrusions réelles
  - Pertinence réelle de l'heuristique (?)
  - Résultats pratiques en ligne
- Fusion de données dans un contexte de flux de données
  - Réduire le nombre de sources pour accélérer le traitement
  - Pb de choix dynamique des sources

# Conclusions

- Problème pratique des flux de données : adaptation en ligne de la détection d'intrusions à partir de logs de serveur Web.
- Solution basée sur :
  - La fusion de décisions fournies par de multiples indices,
  - Une heuristique de confrontation de résultats pour le choix des adaptations
- Implémentation de LogAnalyzer (et outils divers)
- Évaluation avec de la simulation d'intrusions
- Résultats encourageant et perspectives ouvertes
  - Pour la détection d'intrusions
  - Pour l'adaptation dans un contexte de flux de données