Florent Masseglia – INRIA Sophia Antipolis-Méditerranée
Equipe-Projet AXIS

Fouille de données

- Principes généraux du data mining
- Sécurité
- Fouille de flots
- Réseaux de capteurs

http://www.inria.fr/sophia/teams/axis

L’équipe-projet AxIS

- Analyse et Amélioration des Systèmes d’Information Dirigées par l’Usage
- Bi-localisée Sophia – Rocquencourt
- Objectifs 2008-2012 :
  - Analyse d’un SI et ECD
  - Aides pour analyser les réseaux sociaux et améliorer la recherche d’information
  - Elaboration d’une plateforme d’expérimentation FOCUS

L’équipe-projet AxIS

- Analyse d’un SI et ECD :
  - Fouille dans les flots de données (ANR Midas, ARC Sésur)
  - Fouille de données d’un SI (usage, contenu et structure)
  - Semantic Web Mining …
  - Gestion des connaissances en IS mining (domaine de l’analyste)

L’équipe-projet AxIS

- Aides pour analyser les réseaux sociaux et améliorer la recherche d’information :
  - ANR Eiffel
  - Projet Quaero - France-Allemagne
  - ANR Intermed
  - Démarrage d’une thèse en septembre

L’équipe-projet AxIS

- Elaboration d’une plateforme d’expérimentation FOCUS
  - Classes d’applications :
    - Démocratie participative, Développement durable (ANR Intermed)
    - Transport et tourisme (Color INRIA CusCov, ANR Eiffel)
  - Support : une des plateformes du CPER (PACA), dans le cadre du «Living Lab ICT» (ENoLL) de Sophia Antipolis, soumission FP7 infrastructures de recherche

Types d’extraction de connaissance

- Classification

  créé une fonction qui classifie une donnée élémentaire parmi plusieurs classes prédéfinies existantes

- Régression

  créé une fonction qui donne une donnée élémentaire à une variable de prévision avec des données réelles

- Segmentation (clustering, classement)

  rechercher à identifier un ensemble fini de catégories ou groupe en vues de décrire les données

Types d’extraction de connaissance (2)

- Résumé

  affiner une description compacte d’un sous-ensemble de données

- Modelage des dépendances

  trouver un modèle qui décrit des dépendances significatives entre les variables

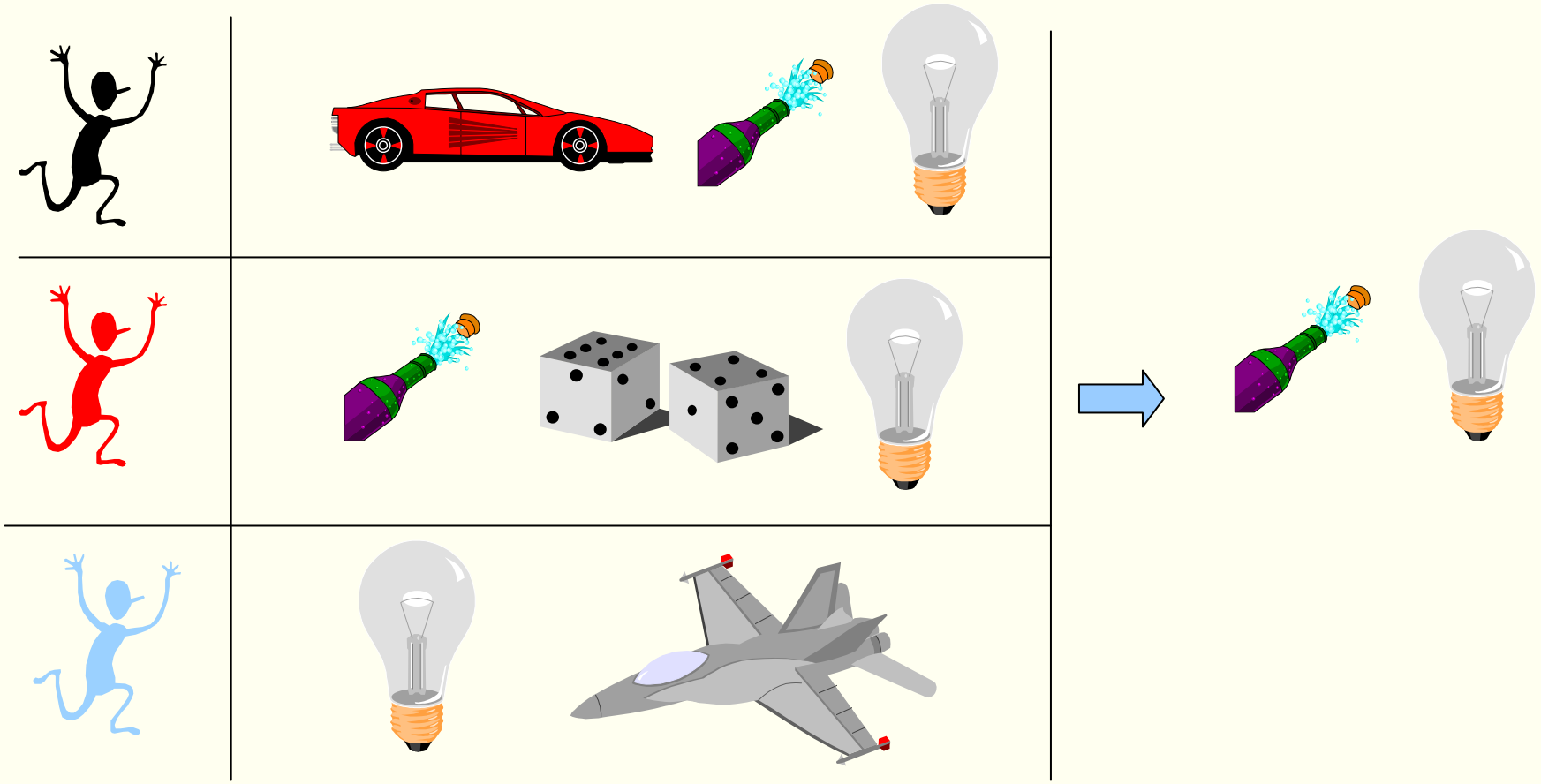
- Détection de changement et déviation

  découvrir les changements les plus significatifs dans les données

Types d’extraction de connaissance (2)

- Extraction de fréquents :
  - Itemsets fréquents (et règles d’association)
    - comportements de clients d’un supermarché
  - Motifs séquentiels fréquents
    - comportements d’utilisateurs d’un site Web
    - comportements de touristes (véhicule + GPS)

Extraction d’itemsets fréquents



Extraction de règles d’association

- 20 % des clients qui achètent du beurre et du pain achètent aussi du lait.
- 15 % des clients qui achètent de la bière et des gâteaux achètent aussi des couches.
- 64 % des étudiants qui suivent le cours “ *Introduction à Unix* ”, suivent également le cours de “ *Programmation C* ” et 34 % de tous les étudiants ont en fait suivis les deux cours.

Extraction de règles d'association

Transaction	Items
10	A, B, C
20	A, C
30	A, D
40	B, E, F

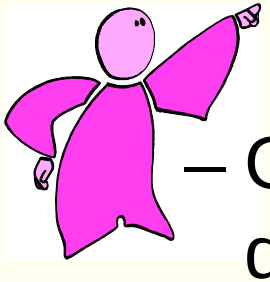
Motifs fréquents	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

$$support\,(A \Rightarrow C) = \frac{support(\{A\} \cup \{C\})}{|D|} = \frac{2}{4} = 50\%$$

$$confiance\,(A \Rightarrow C) = \frac{support(\{A\} \cup \{C\})}{support\,(\{A\})} = \frac{2}{3} = 66.7\%$$

Extraction de règles d’association

Schéma algorithmique de base



- Génération de tous les ensembles fréquents
  - *support minimum*
- Génération à partir des ensembles fréquents de toutes les règles d ’associations
  - *confiance dans la règle*

Extraction de règles d’association

Génération de tous les ensembles fréquents

– Semble facile ...

… problème principal des algorithmes de règles d ’association …

– 1000 items => 2^{1000} ensembles à tester

Extraction de règles d'association

Premiers pas : le principe d'Apriori

- La « référence » dans le domaine
- Equipe d 'IBM Almaden (R. Agrawal et R. Srikant, VLDB94)

- Proposition d'une génération de candidats économe :

  « Tout motif ayant un sous-motif non fréquent n'est pas fréquent »

- Proposition d'une structure efficace

Motifs séquentiels fréquents

- Exploite des données… séquentielles

  Séquence de données : liste ordonnée d'itemsets de la forme

  < itemset(T1) itemset(T2) … itemset(Tn) >

  ex: < (jour1 / caméscope, K7) (jour 3 / batterie) >

- Extrait des motifs fréquents en conservant l'information d'ordonnancement des données en entrée grâce à sa notion d'inclusion pour le support :

  $S_1=<a_1\ a_2 \ldots\ a_n>$ et $S_2=<b_1\ b_2 \ldots\ b_n>$ on a $S_1 \subseteq S_2$ si

  $\exists\ i_1 < i_2 < \ldots < i_n$ / $a_1 \subseteq b_{i1}, \ldots, a_n \subseteq b_{in}$

Extraction de motifs séquentiels

Exemple

	Date 1	Date 2	Date 3	Date 4
C1	10 30 140	20 70 110	40 50	20 60 80
C2	50 120 150	10 30	20 80 140	20 60 70
C3	40 60 70	50	70 80 130	70 90
C4	10 30 70	20 30 150	20 60 110	20 90 130

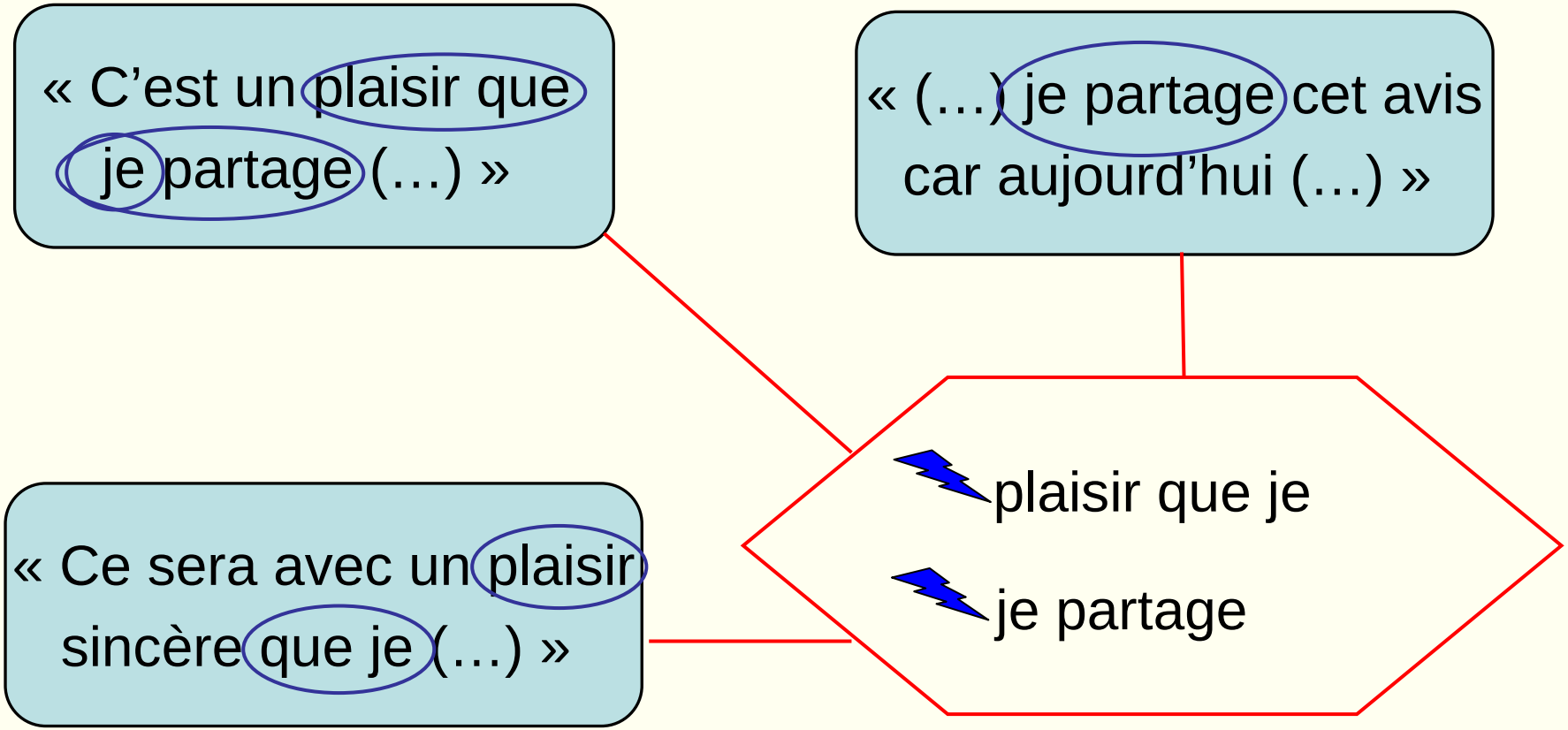
Support minimum = 60% (3 clients suffisent), 3 motifs fréquents :

< (70) >
< (50) (80) >
< (10 30) (20) (20 60) >

Applications des motifs séquentiels

- **Fouille de textes** : extraire des sous-phrases communes à un ensemble de textes.
- Appliqué sur les discours du président de l'assemblée nationale.
- Est-ce que ses discours utilisent les mêmes mots clés, dans le même ordre ?

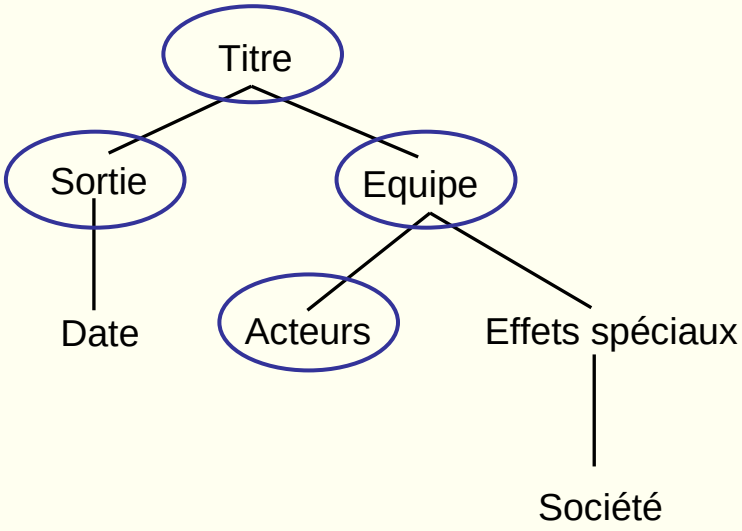
Applications des motifs séquentiels

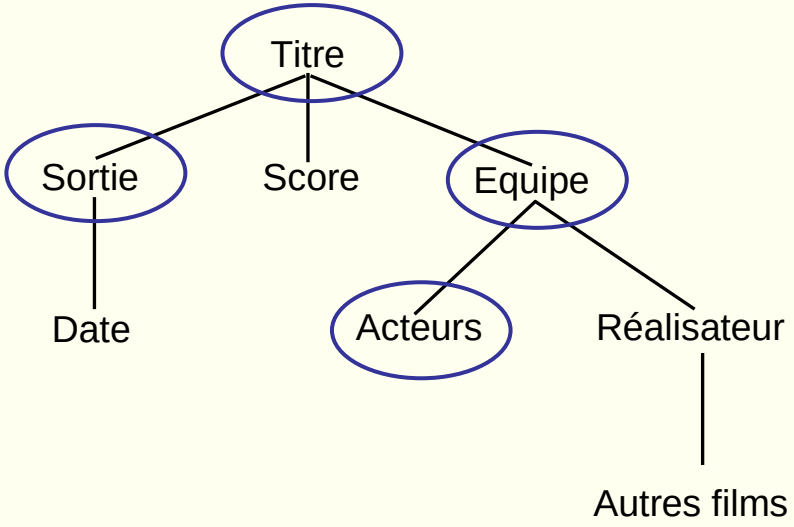


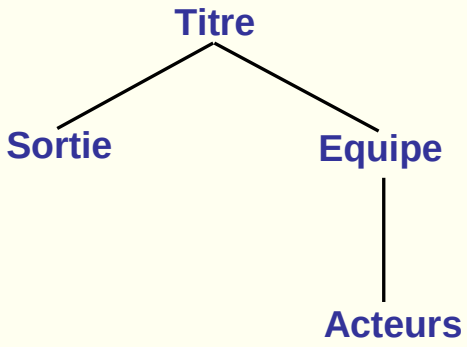
Applications des motifs séquentiels

- **Fouille de graphes** : extraire des sous-graphes communs à un ensemble de graphes.
- Appliqué sur les enregistrements de l'IMDB
- Est-ce que fiches des films/acteurs peuvent être restructurées autour d'un format minimum ?

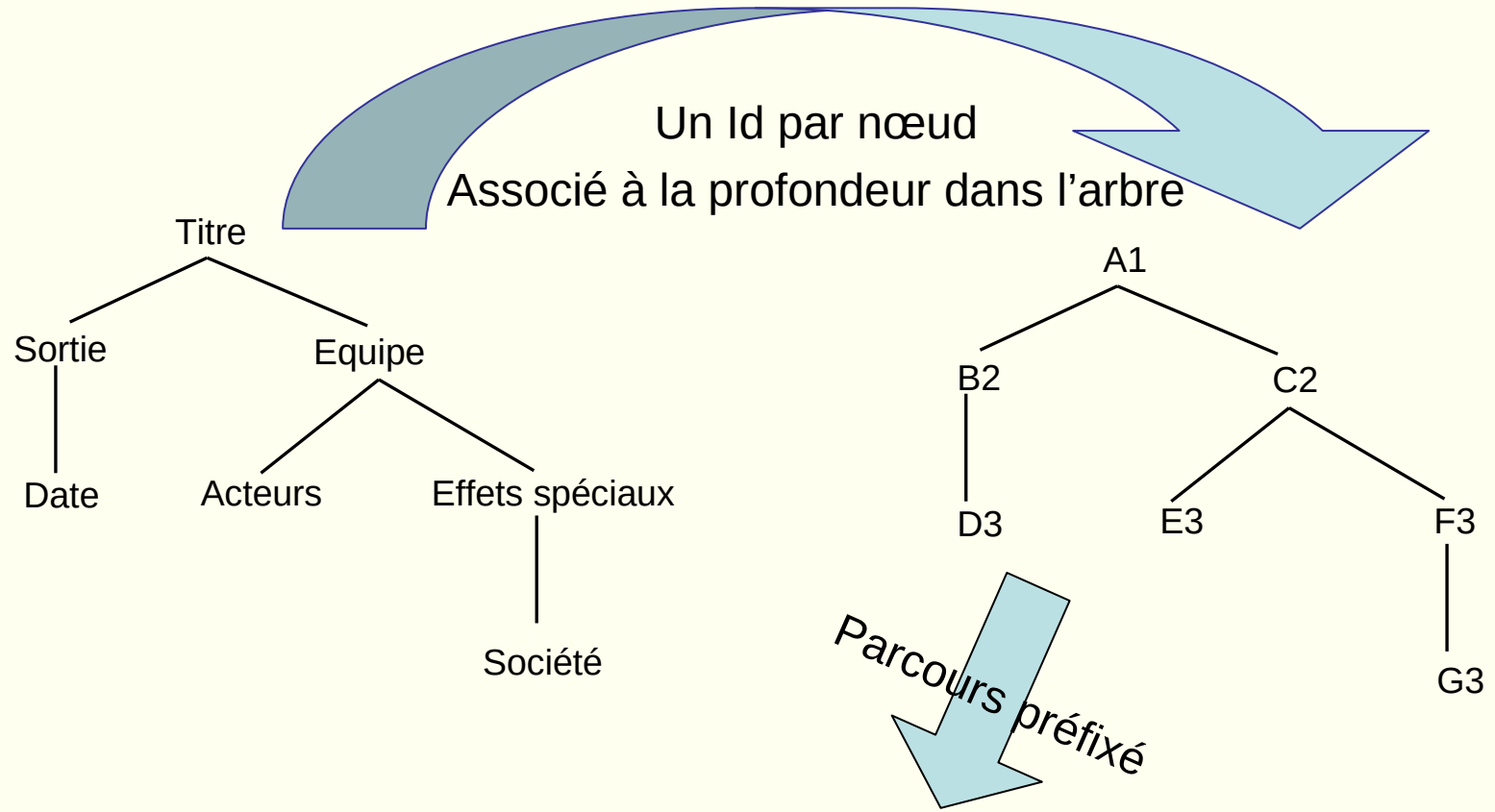
Applications des motifs séquentiels







Applications des motifs séquentiels



Séquence correspondant au schéma :

< (A1) (B2) (D3) (C2) (E3) (F3) (G3) >

Fouille des usages du Web? (rapidement…)

- Les serveurs Web enregistrent les séquences d'accès au site.
- A l'origine le but était de détecter les erreurs (type 404)
- La communauté « fouille de données » est née au début des années 90… (les logs d'accès Web sont rapidement devenus « appétissants »).
- Objectif : extraire tout ce qui se produit avec une fréquence « élevée » (problème classique en fouille de données).

- Exemple (Yahoo News?):
- "7% des utilisateurs naviguent de la manière suivante :
  URL1 = Chiffres_Pouvoir_d_Achat.html ;
  URL2 = Tarifs_Baril_Petrole.html ;
  URL3 = People_Nouvelle_Star.html"

Fouille & sécurité

Le cas des IDS

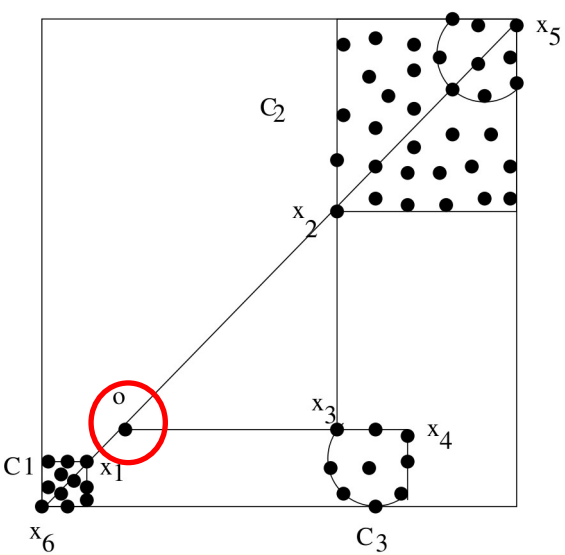
1. Basés sur les signatures :
   - Hyper robuste aux attaques connues
   - Hyper inefficace pour les nouvelles attaques
2. Basés sur les anomalies :
   - Très efficace pour détecter les nouvelles attaques
   - Très efficaces pour saturer l'expert avec un maximum de fausses alertes ! !

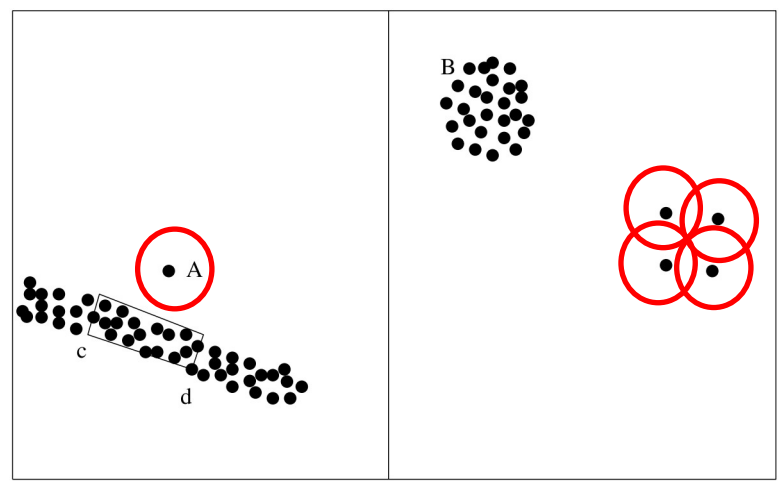
Fouille & sécurité

- Trois thèmes principaux :
  - Clustering & détection d'anomalies.
  - Extraction de motifs off-line pour générer des règles de détection (signatures).
  - Construction d'un modèle du système et comparaison des nouvelles entrées avec le modèle…

Fouille & sécurité

- Clustering et détection d'outliers





Fouille & sécurité

- Clustering et détection d'outliers
  - Nombre élevé de faux positifs
  - Temps de calcul du clustering
  - Découverte en temps réel = capacité à travailler sur les flots de données

Fouille & sécurité

- Fouille de données off-line pour enrichir la base de signatures :
  - Trouve des navigations sur un thème d'une équipe :
    http://www-sop.inria.fr/epidaure/foie3d/ : <(endo4.html) (endo5.html) (endo6.html) (endo8.html) (endo9.html) (endo10.html) (endo11.html) (endo12.html)>
  - Trouve aussi des tentatives d'intrusion :

http://www.inria.fr/ : <(scripts/root.exe) (c:/winnt/system32/cmd.exe) (..%255c../..%255c../winnt/system32/cmd.exe) (..%255c../..%255c/..%c1%1c../..%c1%1c../..%c1%1c../winnt/system32/cmd.exe) (winnt/system32/cmd.exe) (winnt/system32/cmd.exe) (winnt/system32/cmd.exe)>

Fouille & sécurité

- Comparaison on-line au modèle :
  - Thème de l'ARC SéSur (AxIS, Dream, Tatoo, KDD).
  - Travaux de Wei Wang (Post-doc).

- Test sur des données de log HTTP de l'INRIA Sophia
  - Taille des données: 561M
  - Nbre de requêtes : 1,449,379
  - Durée : 3 jours, 2 heures et 10 min

Fouille & sécurité

- Données de log :

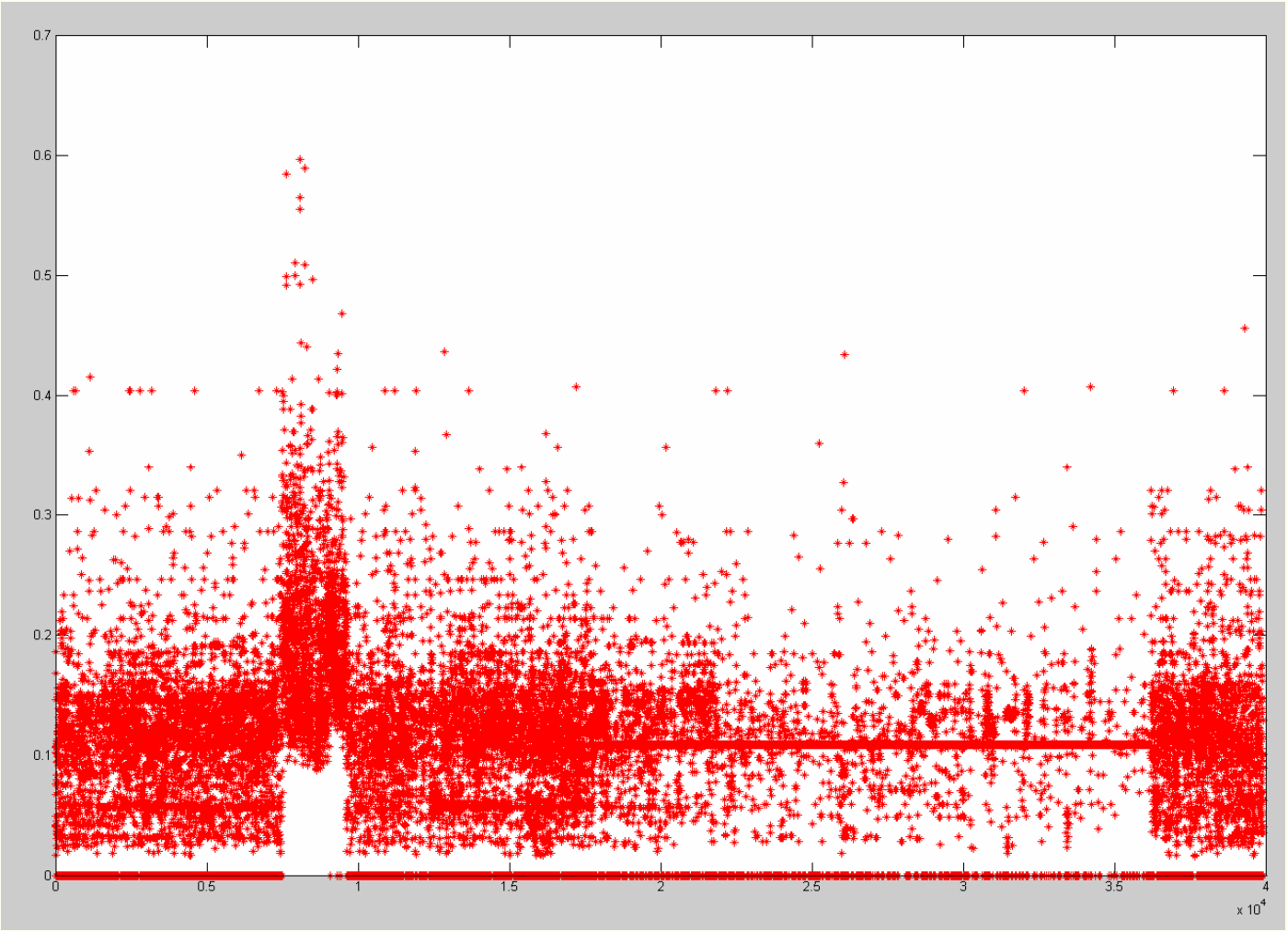
  salmacis.inria.fr - - [10/May/2007:18:27:32 +0200] "GET /cgi-bin/db4web_c/dbdirname//etc/passwd HTTP/1.0" 404 4856 "-" "Mozilla/4.75 (Nikto/1.36)«

- Calculer la distribution des caractères dans l'URL de la requête :
  - Uniquement sur la plage ASCII 33-127
  - Chaque requête est représentée sur un vecteur de taille 95
  - Vecteur après la transformation : 0 0 0 0 0 0 0 0 0 0 0 0 1 0 6 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 2 4 3 4 3 0 1 0 3 0 0 0 1 2 0 1 0 1 2 1 0 0 2 0 0 0 0 0 0 0 0 0
  - La classification en ligne est basée sur ces vecteurs

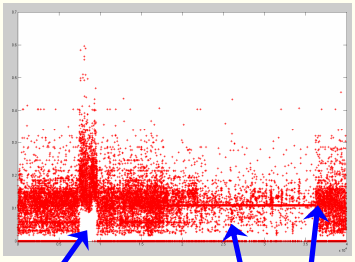
Fouille & sécurité

- Détection d'anomalies
  - Sélectionner les 400 premières requêtes (normales) comme base de référence.
  - Calculer la distance entre les nouvelles entrées et les 400 enregistrements de référence.
  - Sélectionner une distance minimale comme *indice d'anomalie.*

Fouille & sécurité



Fouille & sécurité



- Détection d'anomalies en fonction de la distance
  - Indice=0.14
  - 282/457=61.7% (Taux de détection)
  - 8727/39456=22.1% (Faux positifs)
- Détection des changements de modèle (concept drift)
  - En fonction du taux d'anomalies
  - Indicateurs de type Page-Hinkley
- Work in progress…
  - Améliorer le taux de détection
  - Améliorer le taux de faux positifs
  - Gérer les changements de modèle (historique des concepts drifts, aspects flots)
  - …

Flots de données

Fouille de

flots de

données

Flots de données

- 30 milliards d'emails par jour - 1 milliard de SMS.
- « China's cellular operators estimate Chinese customers will send around 14 billion Lunar New Year text messages on their mobile phones during the week-long holiday ».
- IP Network Traffic: Jusqu'à 1 milliard de paquet par heure et par routeur. Chaque FAI a des centaines de routeurs ! 75000 tuples par seconde !
- AT&T collecte 100 Gb de données réseau par jour.
- NASA EOS (Earth Observation System) : Données d'observation satellites, jusqu'à 350 Gb par jour.
- eBay : en moyenne 1 million de pages consultées par jour.
- Yahoo: 166 millions de visiteurs par jour; 48 Gb de clickstream par heure !

Flots de données

- Caractéristiques des flots de données :
  - De nouveaux éléments générés en permanence (flux potentiellement infini).
  - Les données doivent être traitées aussi vite que possible.
  - Interdiction de bloquer le flux.
  - Un seul coup d’oeil.
  - Restreint par la mémoire disponible.

Application type : le trafic réseaux

- Securité, Fraude, Analyse, Estimation
  - Combien d'octets échangés entre deux IP ?
  - Liste des 100 IP qui consomment le plus.
  - Durée moyenne d'une session IP ?
  - Identifier les sessions qui durent deux fois plus longtemps que la moyenne.
  - Identifier les IP impliquées dans plus de 100 sessions.
  - Quels sont les k éléments les plus demandés ?

    (par exemple les 50 URLs les plus accédées)
  - Détecter une forte augmentation des connexions entrantes en provenance d'un pays sur une courte durée.

Fouille de flots de données

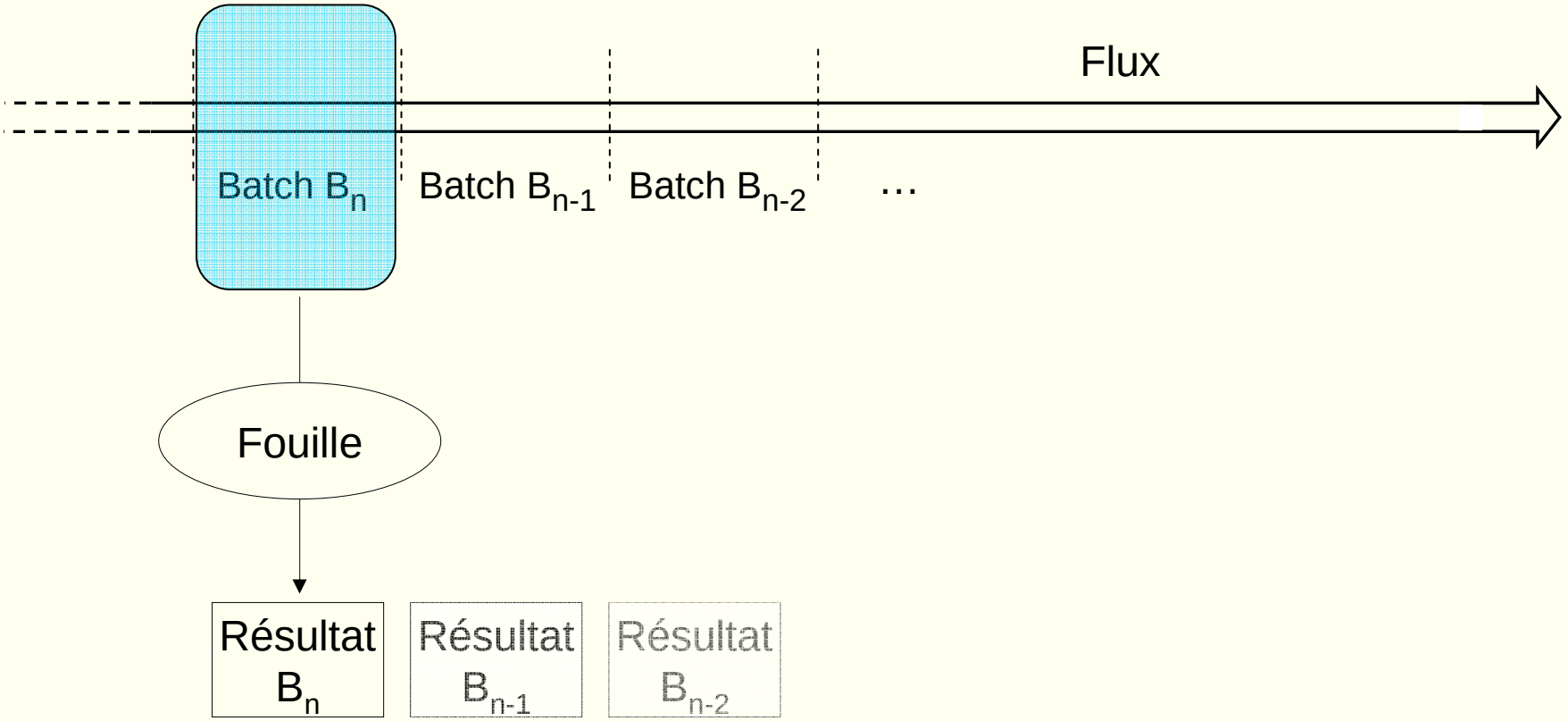
- Irréaliste d'essayer d'extraire précisément les motifs fréquents dans un flot!
- Défi principal : comment extraire les motifs avec une approximation acceptable ?

- Défis associés :
  - Une seule passe: on ne peut pas revenir en arrière !
  - Les fréquents peuvent devenir rares et vice-versa.
  - Besoin de gérer l'historique des connaissances extraites.
  - Besoin de voir les changements importants (ex. intrusions).

Fouille de flux de données

- En raison des caractéristiques/contraintes d'un flux, il est généralement admis qu'une méthode de fouille doit :
  - Sacrifier de son exactitude au profit de sa vitesse d'extraction.
  - Gérer l'historique des « schémas » extraits.

Modèle d’extraction par batch



Gestion de l’historique des connaissances

Approche par « tilted time windows »

Itemset fréquent (a b c)

15 min	15 min	30 min	1 heure	2 heures	…	256 jours
0.17	0.18	0.25	0.12	0.05	…	0

Fouille de flux de données

Aligner les séquences d'un cluster

< (a) (b) (d) >
< (a) (c) (d) >

$\Rightarrow$

< (a) (b) (d) >
< (a) (c) (d) >

< (a:2) (b:1, c:1) (d:2) >

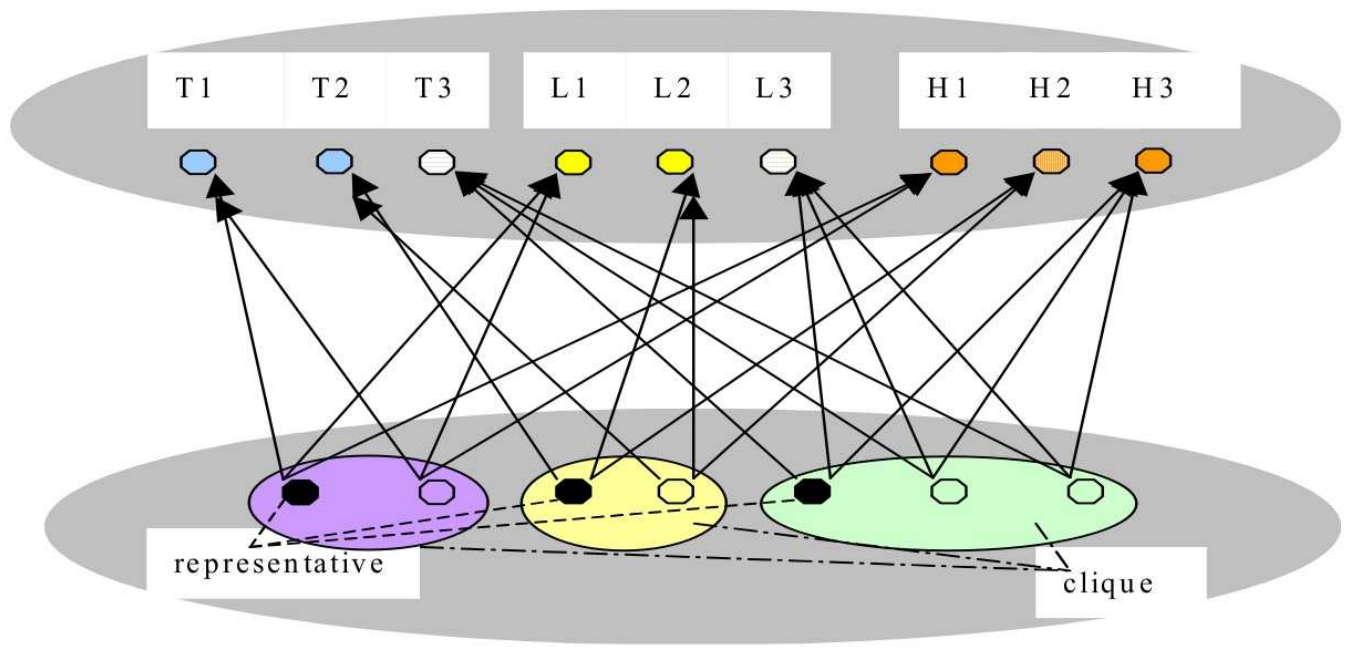
Filtre *k* =1: < (a:2) (b:1, c:1) (d:2) >
Filtre *k* =2: < (a:2) (d:2) >

Fouille de données et réseaux de capteurs
(petit tour d'horizon « biaisé »)

- Fouille de données afin d'améliorer le réseau.
  - Clustering
  - Règles de communication
  - Économies de transmissions
  - Économies de batterie
- Challenge du data mining sur un environnement sur-contraint
  - Distribution du processus de fouille
  - Faibles ressources disponibles
- Applications ?

Fouille de données afin d'améliorer le réseau ?

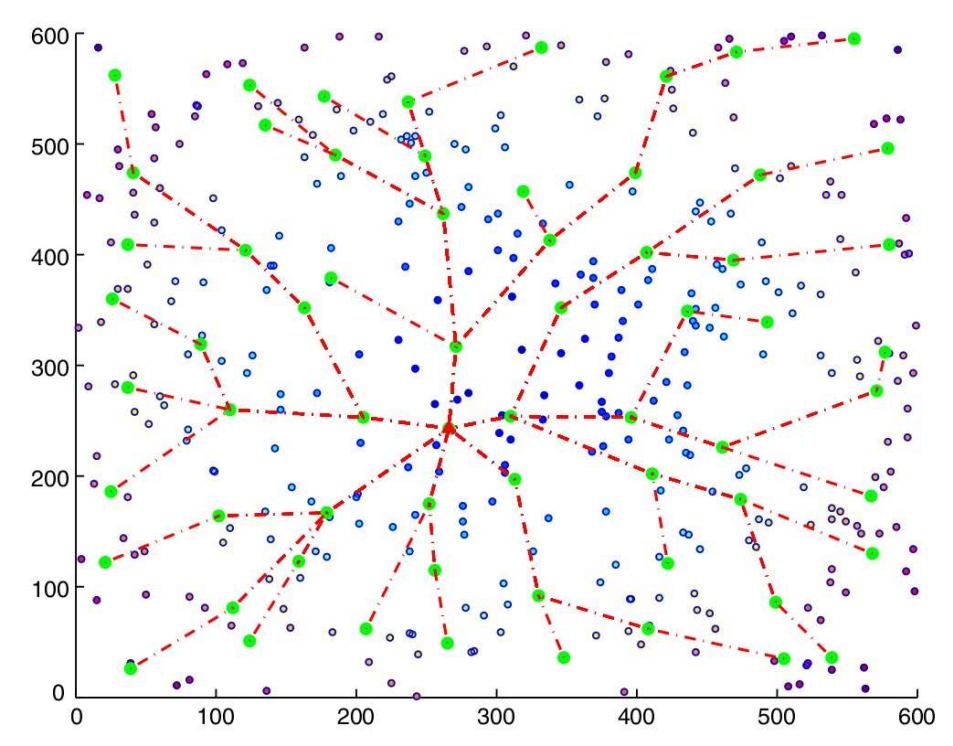
- Clustering des capteurs :
  - Grouper sous un leader afin d'économiser les échanges.



- En particulier pour les nœuds mobiles.

Fouille de données afin d'améliorer le réseau ?

- Clustering des capteurs : « mining sensor energy data »
  - Exploiter les données d'énergie afin de grouper les nœuds.
  - Ne repose pas sur des données de localisation des nœuds.

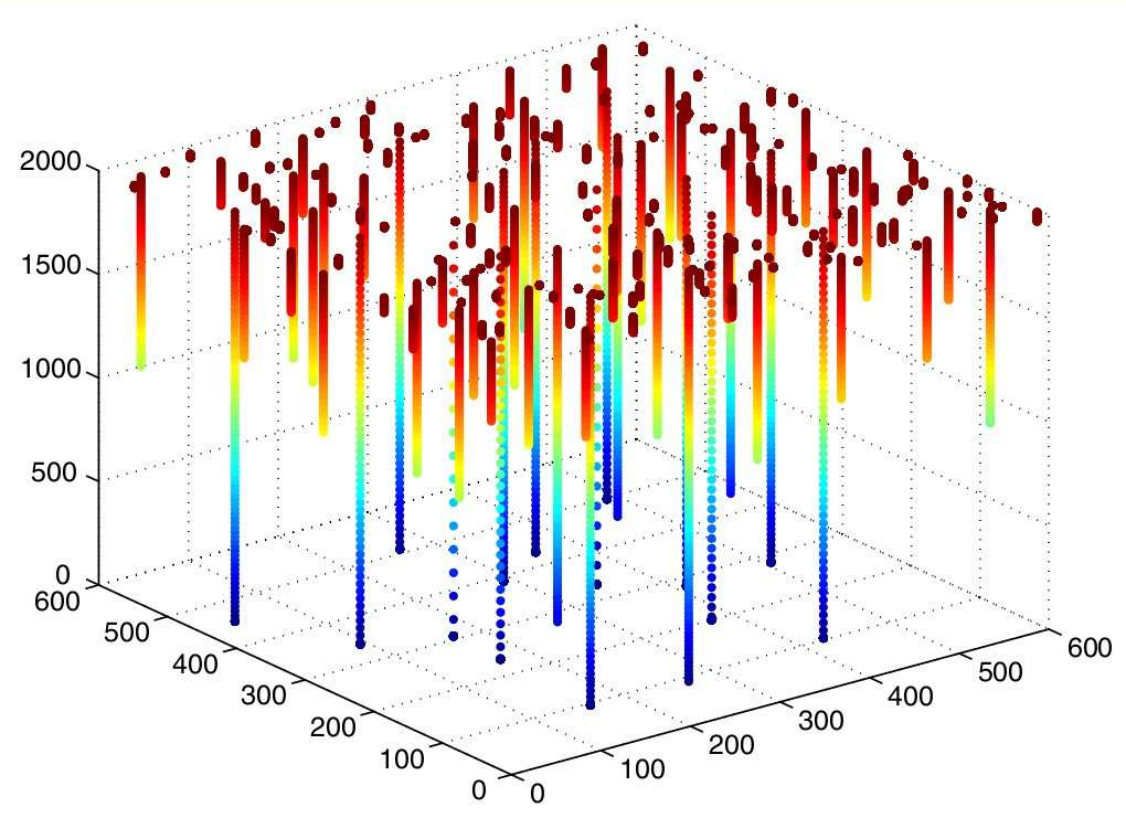


Fouille de données afin d'améliorer le réseau ?

- Clustering des capteurs : « mining sensor energy data »

  - Exploite l'historique de la consommation d'énergie des capteurs

  - Groupe les capteurs en fonction des similitudes de consommations (suppose que les leaders consomment de façon similaire)



Fouille de données afin d'améliorer le réseau ?

- Règles d'association pour optimiser les communications

  (Boukerche and Samarah, Ottawa, IEEE T Parallel and Distrib. Syst.)

- Extraire des règles basées sur des activités communes

- Exemple : ($s_1 s_2 \Rightarrow s_3$, 90%, t) « après des transmissions de s_1 et s_2 il y a 90% de chances d'observer une transmission de s_3 dans un délai de t unités de temps ».

- Implique de découvrir le motif (itemset) fréquent ($s_1 s_2 s_3$) et de calculer support, confiance et délais.

Fouille de données afin d'améliorer le réseau ?

- ($s_1 s_2 \Rightarrow s_3$, 90%, t)
- Objectifs applicatifs :
  - Surveiller les comportements des nœuds (on attend un événement de s_3 à 90% et il ne se produit pas dans le délai t).
  - Participer aux efforts de gestion de ressources (on peut mettre un nœud en stand by car il ne communiquera certainement pas dans les *n* prochaines unités de temps).
  - Prédire les comportements des nœuds (source du prochain événement)

Fouille de données afin d'améliorer le réseau ?

- Règles d'association pour optimiser les communications

  (Gaber et al., ACM SAC'08)
- Approche précédente basée sur les méta données (« un événement s'est produit »).
- Cette approche se base sur le contenu des événements (« le nœud s_1 a envoyé les information *xy* »).
- Un ensemble de capteurs S_1, S_2, …. S_n et un ensemble de valeurs de son *s*, de température *t* et de lumière *l*.
- Les valeurs de lumière pour S_1 et S_2 peuvent être similaires (placés dans la même zone).
- Discrétisation des valeurs : [0,299] ='L', [300,699]='M' et [700,1000]='H'
- Transactions de type {S_n; t_n, s_n, l_n, time}

Fouille de données afin d'améliorer le réseau ?

- Règles d'association pour optimiser les communications
- Transactions de type $\{S_n; t_n, s_n, l_n, time\}$
- Exemples de règles extraites :
  - S0L[H] ⇨ S1L[H], 100%
  - S0L[L] ⇨ S0T[M], 85.7%
- Objectif : éviter des communications si on peut prévoir les valeurs à transmettre.

Exploitation des données du réseau

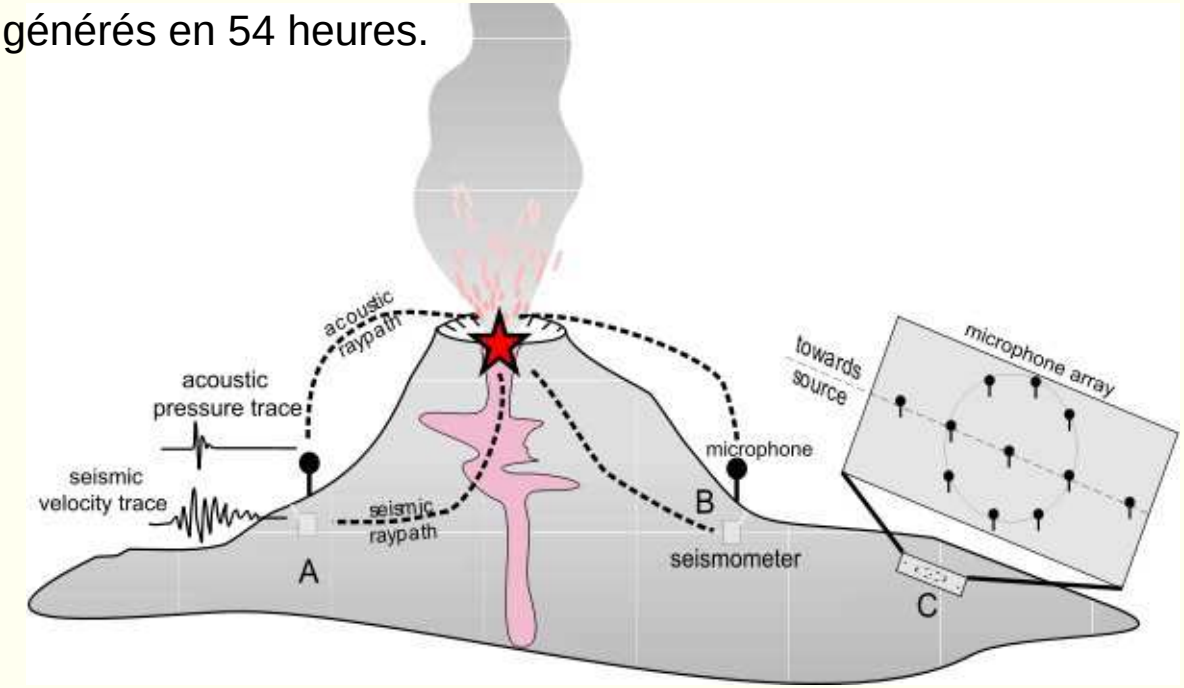
- Prévision de charge dans un réseau électrique :
  (Rodriguez & Gama, Porto)

- Capteurs distribués dans le réseau électrique.
- Apprentissage en ligne et détection de changement.
- Objectif : maintenir en temps réel un modèle de clustering du réseau et un modèle prédictif capable de détecter les changement.

Exploitation des données du réseau

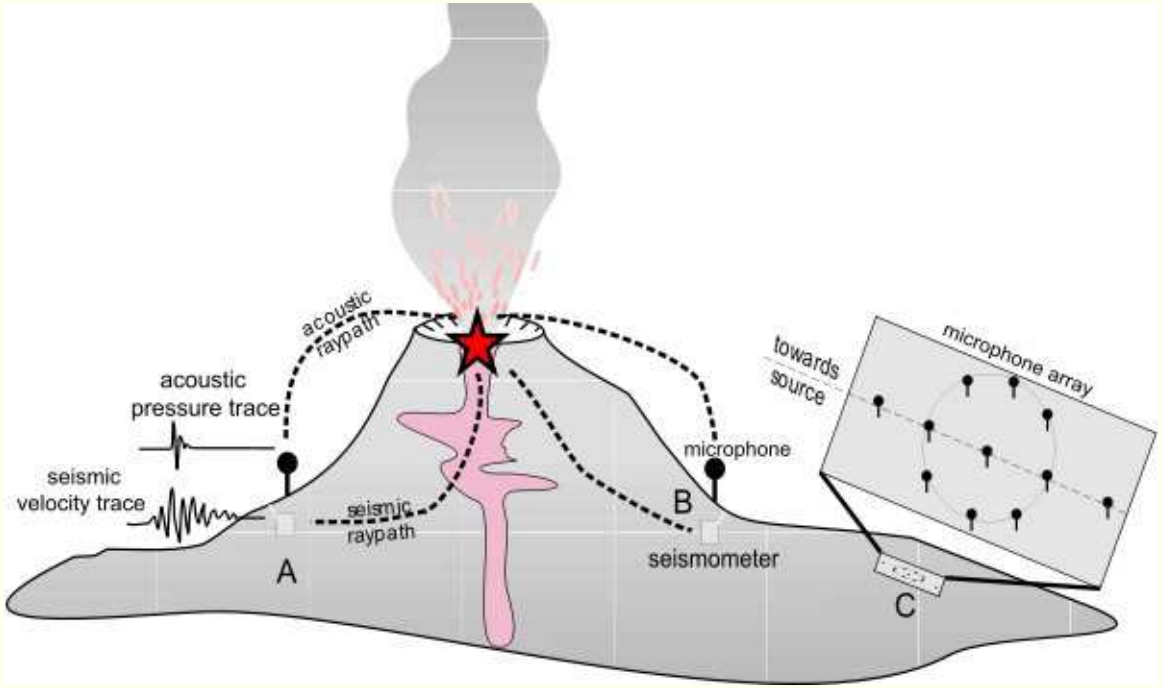
- Capteurs autour d'un volcan :

  (Werner-Allen, Johnson, Ruiz, Lees and Welsh, Harvard)
- Capteurs (distribués géographiquement autour du volcan Tungurahua, équateur).
- 1,57 Gb de signaux générés en 54 heures.



Exploitation des données du réseau

- « 20 minutes après un événement d'éruption, des chocs sismiques fréquents se produisent. »
- « Autour d'un événement d'émission de gaz, des événements sismiques sont probables. »



Exploitation des données du réseau

- Monitoring d'un réseau de télécommunications

- Capteur mesurant l'utilisation de la bande passante sur des liens.
- Détection de congestion et de comportements anormaux.

- « 30% des liens sur un switch sont souvent fortement sollicités au même moment et un ralentissement est observé à ce moment ».

- « Le trafic sortant d'un sous-réseau est nettement plus élevé que le trafic entrant ».

Exploitation des données du réseau

- Surveiller des Pandas en Chine

  (Ma et al., Beijing)

- Capteurs météo autour de l'habitat du panda.
- Capteurs sur le panda (Huanhuan).

- « Est-ce que Huanhuan a des symptômes anormaux par rapport au passé ? Est-ce que d'autres pandas ont des symptômes similaires ? »

- « Y a-t-il des corrélations entre les attributs des pandas et ceux de leur habitat ? ».

Bilan sur les capteurs

- Techniques de fouille sur les réseaux de capteurs
- Pour améliorer le réseau ou pour exploiter ses données…
- Centralisées :
  - Ne passent pas à l'échelle
  - Consomment de l'énergie et de la bande passante (fouille des données transmises à un sink)
- Distribuées :
  - Limitées par les capacités embarquées du capteur
- Challenges « rassurants » (en data mining) :
  - Aspects flots (ressources limitées)
  - Distribuer les calculs
  - Concept drift