

# Intelligence Artificielle et Fouille de données

questions ouvertes et perspectives pour les thématiques des JFPDA

**Thomas Guyet** avec des idées de  
René Quiniou et Alexandre Termier

Agrocampus-Ouest/Inria/IRISA

JFPDA, July 2015, Rennes

# Objectifs de la présentation

- Présentation des thématiques de recherche locales
  - ⇒ équipe DREAM en transition vers une équipe plus centré sur les thèmes de la fouille de données
- Objectif ambitieux de pointer des rapprochements entre
  - les thématiques des JFPDA (décision, planification, méthodes formelles, etc) **dont je connais pas grand chose !**
  - les thématiques de *pattern mining* **qui m'occupe la plupart de mon temps**
- ⇒ Les questions sont principalement portées dans le sens de l'intérêt des approches d'intelligence artificielle pour améliorer les méthodes de fouille de données
- ⇒ Deux objectifs auxquelles elles peuvent contribuer
  - extraire de "meilleurs" motifs en intégrant du raisonnement et des connaissances
  - améliorer l'utilisation des motifs extraits en les transformant en connaissances "utilisables"

# Outline

- 1 Monitoring dynamical systems and data mining in Rennes
  - DREAM Team
  - Monitoring dynamical systems
  - Quelques applications
  - Réorganisation des objectifs de recherche
- 2 Une approche exemple de Pattern Mining et IA : extraction de motifs séquentiels avec ASP
  - Pattern mining
  - Fouille avec ASP
  - Cas exemple : analyse de trace de patients
- 3 Pattern mining et Intelligence Artificielle : Questions ouvertes
  - Planification et extraction de motifs
  - Composer des chaînes d'extraction de connaissances
  - Raisonnement (décision/argumentation) sur et avec les motifs

# DREAM Team

A team working on artificial intelligence tools to bring novel tools to diagnose, to monitor and to model dynamic systems.

- D : Diagnosis
  - ⇒ causal reasoning, logic programming, ASP
- RE : Recommending
  - A : Action
    - ⇒ argumentation, logic programming
- M : Modelling
  - ⇒ automata, model checking approaches,
  - ⇒ data mining, machine learning, stream mining, spatial and temporal

# Monitoring dynamical systems

## Dynamical system

- A **dynamical system** is a system that evolves in time,
- The evolution of the system depends on the **context**, the context may change over the time,
- Some systems **behaviours** are *faulty*,
- A dynamical system generates traces that inform about the system state along time.

## Monitoring/diagnosing dynamical systems

- Monitoring a dynamical system is deciding, at each instant, if the dynamical is (or will be) in a faulty behaviour or not.
- Diagnosing is identifying the explanation of a (faulty) behaviour

# Some system examples

- living systems : patients in ICU<sup>1</sup>, cows, ...
  - ⇒ monitoring the physiologic signals of the patient.
  - ⇒ early detection of major events (vélagés, chaleurs).
- environmental systems : vegetation, watershed ...
  - ⇒ monitoring the evolution of the environment from satellite images
- technical systems : engines, electronic chips, electrical networks, ...
  - ⇒ prediction of regional electrical consumption from individual instantaneous consumption records
  - ⇒ detection of anomalies in CHIPSET execution traces (Go/minutes of data)
- digital systems : software, web server, ...
  - ⇒ detection of intrusions in web servers from logs

---

## 1. Intensive Care Units

# Model-based diagnosis/monitoring

## Definition

Model-based diagnosis uses explicit models of the system to be diagnosed.

- monitoring is comparing the current state to the models,
  - model(s) of faulty behaviours : require to know the faulty behaviours (!), very sensitive to context changes,
  - model(s) of normal behaviours : difficult to construct, generate more false negatives,
- reasoning can use general knowledge about the system,

## Building the model

- The models are difficult to construct
- The richer is the model, the more it is difficult to learn
- The context changes require on-line model building

# Using trace patterns as model

- The system execution trace reveals the state of the system : we assume that traces holds enough information to discriminate *normal* and *faulty*
- Execution traces : numerical/symbolic, regular sampling or event based,
- Patterns in traces are interesting to characterize the system behaviours.
  - 1 patterns are mapped to behaviours
  - 2 a set of patterns is mapped to a behaviour (more robust)

## Traces as streams of itemsets

Traces are timestamped discrete events : stream of itemsets

- numerical signals have to be discretized
- timestamped are discrete values
- several events may append at the same time



# Using trace patterns as model : two classical tasks

## Monitoring the system with patterns

- The objective is to use "patterns" (or an formal object constructed from the patterns) to decide about the state of the system
- Current difficulties :
  - compare the current state of the system with the patterns
  - concept drift : the (normal/faulty) behavior of the system may change over time
  - efficiently extracting patterns on the fly (datastream mining)

## Supporting the expert to interpret data

- The objective is to use the "patterns" to give an abstract view of the system behaviors and to support the analysis of the data.
- The expert will then create appropriate models to
- difficulties
  - being able to process more and more data
  - being able to extract more accurate patterns
  - being able to integrate the expert knowledge in the mining process

# Example : Electrocardiogram (ECG)

## Objective

Extracting patterns that are characteristic to some cardiac diseases.

## Electrocardiogram segmentation

- electrocardiogram may be abstract by a sequence of events (three typical events : P, QRS, T),
- signal processing may automatically annotate time series.

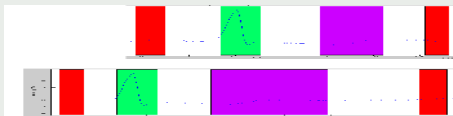


Figure : Two examples of ECG : normal (left), disease (right)

- monitoring usage : detecting on-line the first pattern triggers an alarm
- analyzing usage : automatically annotate ECG

# Spatial data mining

## Spatial data mining

Objectives : support the understanding of the impact of spatial organizations on agro-environmental processes.

- mining attributed graphs for characterizing and simulating agricultural landscapes.
- discriminant learning of spatial organization (PayTal), water and pesticide runoffs prediction (SACADEAU), urban sprawling prediction (PAYTAL)
- simulation of "realistic" landscape :  
mixing patterns and expert knowledge

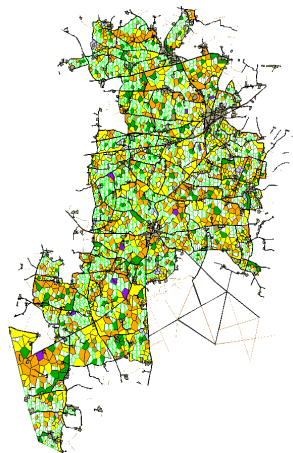


Figure : Simulated landscape from frequent spatial patterns

# Temporal data mining

## Temporal data mining

Objectives : modelling the behaviours of dynamical systems for online monitoring

- Data : time series, sequential data
- Modelling objective : building model for dynamical system monitoring
- Some challenges :
  - extracting multi-scale symbolic patterns [MGQ13]
  - extracting frequent behaviours from traces
  - online adaptation to concept-changes

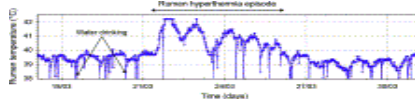


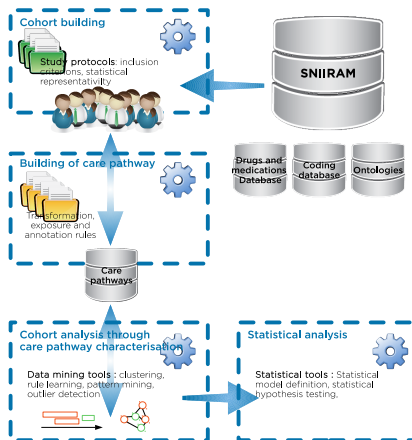
Figure : Temperature of cow : time series analysis

```

13/10/2013 10:00:00 [1] 32.5
13/10/2013 10:05:00 [1] 32.8
13/10/2013 10:10:00 [1] 33.1
13/10/2013 10:15:00 [1] 33.4
13/10/2013 10:20:00 [1] 33.7
13/10/2013 10:25:00 [1] 34.0
13/10/2013 10:30:00 [1] 34.3
13/10/2013 10:35:00 [1] 34.6
13/10/2013 10:40:00 [1] 34.9
13/10/2013 10:45:00 [1] 35.2
13/10/2013 10:50:00 [1] 35.5
13/10/2013 10:55:00 [1] 35.8
13/10/2013 11:00:00 [1] 36.1
13/10/2013 11:05:00 [1] 36.4
13/10/2013 11:10:00 [1] 36.7
13/10/2013 11:15:00 [1] 37.0
13/10/2013 11:20:00 [1] 37.3
13/10/2013 11:25:00 [1] 37.6
13/10/2013 11:30:00 [1] 37.9
13/10/2013 11:35:00 [1] 38.2
13/10/2013 11:40:00 [1] 38.5
13/10/2013 11:45:00 [1] 38.8
13/10/2013 11:50:00 [1] 39.1
13/10/2013 11:55:00 [1] 39.4
13/10/2013 12:00:00 [1] 39.7
13/10/2013 12:05:00 [1] 40.0
13/10/2013 12:10:00 [1] 40.3
13/10/2013 12:15:00 [1] 40.6
13/10/2013 12:20:00 [1] 40.9
13/10/2013 12:25:00 [1] 41.2
13/10/2013 12:30:00 [1] 41.5
13/10/2013 12:35:00 [1] 41.8
13/10/2013 12:40:00 [1] 42.1
13/10/2013 12:45:00 [1] 42.4
13/10/2013 12:50:00 [1] 42.7
13/10/2013 12:55:00 [1] 43.0
13/10/2013 13:00:00 [1] 43.3
13/10/2013 13:05:00 [1] 43.6
13/10/2013 13:10:00 [1] 43.9
13/10/2013 13:15:00 [1] 44.2
13/10/2013 13:20:00 [1] 44.5
13/10/2013 13:25:00 [1] 44.8
13/10/2013 13:30:00 [1] 45.1
13/10/2013 13:35:00 [1] 45.4
13/10/2013 13:40:00 [1] 45.7
13/10/2013 13:45:00 [1] 46.0
13/10/2013 13:50:00 [1] 46.3
13/10/2013 13:55:00 [1] 46.6
13/10/2013 14:00:00 [1] 46.9
13/10/2013 14:05:00 [1] 47.2
13/10/2013 14:10:00 [1] 47.5
13/10/2013 14:15:00 [1] 47.8
13/10/2013 14:20:00 [1] 48.1
13/10/2013 14:25:00 [1] 48.4
13/10/2013 14:30:00 [1] 48.7
13/10/2013 14:35:00 [1] 49.0
13/10/2013 14:40:00 [1] 49.3
13/10/2013 14:45:00 [1] 49.6
13/10/2013 14:50:00 [1] 49.9
13/10/2013 14:55:00 [1] 50.2
13/10/2013 15:00:00 [1] 50.5
13/10/2013 15:05:00 [1] 50.8
13/10/2013 15:10:00 [1] 51.1
13/10/2013 15:15:00 [1] 51.4
13/10/2013 15:20:00 [1] 51.7
13/10/2013 15:25:00 [1] 52.0
13/10/2013 15:30:00 [1] 52.3
13/10/2013 15:35:00 [1] 52.6
13/10/2013 15:40:00 [1] 52.9
13/10/2013 15:45:00 [1] 53.2
13/10/2013 15:50:00 [1] 53.5
13/10/2013 15:55:00 [1] 53.8
13/10/2013 16:00:00 [1] 54.1
13/10/2013 16:05:00 [1] 54.4
13/10/2013 16:10:00 [1] 54.7
13/10/2013 16:15:00 [1] 55.0
13/10/2013 16:20:00 [1] 55.3
13/10/2013 16:25:00 [1] 55.6
13/10/2013 16:30:00 [1] 55.9
13/10/2013 16:35:00 [1] 56.2
13/10/2013 16:40:00 [1] 56.5
13/10/2013 16:45:00 [1] 56.8
13/10/2013 16:50:00 [1] 57.1
13/10/2013 16:55:00 [1] 57.4
13/10/2013 17:00:00 [1] 57.7
13/10/2013 17:05:00 [1] 58.0
13/10/2013 17:10:00 [1] 58.3
13/10/2013 17:15:00 [1] 58.6
13/10/2013 17:20:00 [1] 58.9
13/10/2013 17:25:00 [1] 59.2
13/10/2013 17:30:00 [1] 59.5
13/10/2013 17:35:00 [1] 59.8
13/10/2013 17:40:00 [1] 60.1
13/10/2013 17:45:00 [1] 60.4
13/10/2013 17:50:00 [1] 60.7
13/10/2013 17:55:00 [1] 61.0
13/10/2013 18:00:00 [1] 61.3
13/10/2013 18:05:00 [1] 61.6
13/10/2013 18:10:00 [1] 61.9
13/10/2013 18:15:00 [1] 62.2
13/10/2013 18:20:00 [1] 62.5
13/10/2013 18:25:00 [1] 62.8
13/10/2013 18:30:00 [1] 63.1
13/10/2013 18:35:00 [1] 63.4
13/10/2013 18:40:00 [1] 63.7
13/10/2013 18:45:00 [1] 64.0
13/10/2013 18:50:00 [1] 64.3
13/10/2013 18:55:00 [1] 64.6
13/10/2013 19:00:00 [1] 64.9
13/10/2013 19:05:00 [1] 65.2
13/10/2013 19:10:00 [1] 65.5
13/10/2013 19:15:00 [1] 65.8
13/10/2013 19:20:00 [1] 66.1
13/10/2013 19:25:00 [1] 66.4
13/10/2013 19:30:00 [1] 66.7
13/10/2013 19:35:00 [1] 67.0
13/10/2013 19:40:00 [1] 67.3
13/10/2013 19:45:00 [1] 67.6
13/10/2013 19:50:00 [1] 67.9
13/10/2013 19:55:00 [1] 68.2
13/10/2013 20:00:00 [1] 68.5
13/10/2013 20:05:00 [1] 68.8
13/10/2013 20:10:00 [1] 69.1
13/10/2013 20:15:00 [1] 69.4
13/10/2013 20:20:00 [1] 69.7
13/10/2013 20:25:00 [1] 70.0
13/10/2013 20:30:00 [1] 70.3
13/10/2013 20:35:00 [1] 70.6
13/10/2013 20:40:00 [1] 70.9
13/10/2013 20:45:00 [1] 71.2
13/10/2013 20:50:00 [1] 71.5
13/10/2013 20:55:00 [1] 71.8
13/10/2013 21:00:00 [1] 72.1
13/10/2013 21:05:00 [1] 72.4
13/10/2013 21:10:00 [1] 72.7
13/10/2013 21:15:00 [1] 73.0
13/10/2013 21:20:00 [1] 73.3
13/10/2013 21:25:00 [1] 73.6
13/10/2013 21:30:00 [1] 73.9
13/10/2013 21:35:00 [1] 74.2
13/10/2013 21:40:00 [1] 74.5
13/10/2013 21:45:00 [1] 74.8
13/10/2013 21:50:00 [1] 75.1
13/10/2013 21:55:00 [1] 75.4
13/10/2013 22:00:00 [1] 75.7
13/10/2013 22:05:00 [1] 76.0
13/10/2013 22:10:00 [1] 76.3
13/10/2013 22:15:00 [1] 76.6
13/10/2013 22:20:00 [1] 76.9
13/10/2013 22:25:00 [1] 77.2
13/10/2013 22:30:00 [1] 77.5
13/10/2013 22:35:00 [1] 77.8
13/10/2013 22:40:00 [1] 78.1
13/10/2013 22:45:00 [1] 78.4
13/10/2013 22:50:00 [1] 78.7
13/10/2013 22:55:00 [1] 79.0
13/10/2013 23:00:00 [1] 79.3
13/10/2013 23:05:00 [1] 79.6
13/10/2013 23:10:00 [1] 79.9
13/10/2013 23:15:00 [1] 80.2
13/10/2013 23:20:00 [1] 80.5
13/10/2013 23:25:00 [1] 80.8
13/10/2013 23:30:00 [1] 81.1
13/10/2013 23:35:00 [1] 81.4
13/10/2013 23:40:00 [1] 81.7
13/10/2013 23:45:00 [1] 82.0
13/10/2013 23:50:00 [1] 82.3
13/10/2013 23:55:00 [1] 82.6
13/10/2013 00:00:00 [1] 82.9
13/10/2013 00:05:00 [1] 83.2
13/10/2013 00:10:00 [1] 83.5
13/10/2013 00:15:00 [1] 83.8
13/10/2013 00:20:00 [1] 84.1
13/10/2013 00:25:00 [1] 84.4
13/10/2013 00:30:00 [1] 84.7
13/10/2013 00:35:00 [1] 85.0
13/10/2013 00:40:00 [1] 85.3
13/10/2013 00:45:00 [1] 85.6
13/10/2013 00:50:00 [1] 85.9
13/10/2013 00:55:00 [1] 86.2
13/10/2013 01:00:00 [1] 86.5
13/10/2013 01:05:00 [1] 86.8
13/10/2013 01:10:00 [1] 87.1
13/10/2013 01:15:00 [1] 87.4
13/10/2013 01:20:00 [1] 87.7
13/10/2013 01:25:00 [1] 88.0
13/10/2013 01:30:00 [1] 88.3
13/10/2013 01:35:00 [1] 88.6
13/10/2013 01:40:00 [1] 88.9
13/10/2013 01:45:00 [1] 89.2
13/10/2013 01:50:00 [1] 89.5
13/10/2013 01:55:00 [1] 89.8
13/10/2013 02:00:00 [1] 90.1
13/10/2013 02:05:00 [1] 90.4
13/10/2013 02:10:00 [1] 90.7
13/10/2013 02:15:00 [1] 91.0
13/10/2013 02:20:00 [1] 91.3
13/10/2013 02:25:00 [1] 91.6
13/10/2013 02:30:00 [1] 91.9
13/10/2013 02:35:00 [1] 92.2
13/10/2013 02:40:00 [1] 92.5
13/10/2013 02:45:00 [1] 92.8
13/10/2013 02:50:00 [1] 93.1
13/10/2013 02:55:00 [1] 93.4
13/10/2013 03:00:00 [1] 93.7
13/10/2013 03:05:00 [1] 94.0
13/10/2013 03:10:00 [1] 94.3
13/10/2013 03:15:00 [1] 94.6
13/10/2013 03:20:00 [1] 94.9
13/10/2013 03:25:00 [1] 95.2
13/10/2013 03:30:00 [1] 95.5
13/10/2013 03:35:00 [1] 95.8
13/10/2013 03:40:00 [1] 96.1
13/10/2013 03:45:00 [1] 96.4
13/10/2013 03:50:00 [1] 96.7
13/10/2013 03:55:00 [1] 97.0
13/10/2013 04:00:00 [1] 97.3
13/10/2013 04:05:00 [1] 97.6
13/10/2013 04:10:00 [1] 97.9
13/10/2013 04:15:00 [1] 98.2
13/10/2013 04:20:00 [1] 98.5
13/10/2013 04:25:00 [1] 98.8
13/10/2013 04:30:00 [1] 99.1
13/10/2013 04:35:00 [1] 99.4
13/10/2013 04:40:00 [1] 99.7
13/10/2013 04:45:00 [1] 100.0
13/10/2013 04:50:00 [1] 100.3
13/10/2013 04:55:00 [1] 100.6
13/10/2013 05:00:00 [1] 100.9
13/10/2013 05:05:00 [1] 101.2
13/10/2013 05:10:00 [1] 101.5
13/10/2013 05:15:00 [1] 101.8
13/10/2013 05:20:00 [1] 102.1
13/10/2013 05:25:00 [1] 102.4
13/10/2013 05:30:00 [1] 102.7
13/10/2013 05:35:00 [1] 103.0
13/10/2013 05:40:00 [1] 103.3
13/10/2013 05:45:00 [1] 103.6
13/10/2013 05:50:00 [1] 103.9
13/10/2013 05:55:00 [1] 104.2
13/10/2013 06:00:00 [1] 104.5
13/10/2013 06:05:00 [1] 104.8
13/10/2013 06:10:00 [1] 105.1
13/10/2013 06:15:00 [1] 105.4
13/10/2013 06:20:00 [1] 105.7
13/10/2013 06:25:00 [1] 106.0
13/10/2013 06:30:00 [1] 106.3
13/10/2013 06:35:00 [1] 106.6
13/10/2013 06:40:00 [1] 106.9
13/10/2013 06:45:00 [1] 107.2
13/10/2013 06:50:00 [1] 107.5
13/10/2013 06:55:00 [1] 107.8
13/10/2013 07:00:00 [1] 108.1
13/10/2013 07:05:00 [1] 108.4
13/10/2013 07:10:00 [1] 108.7
13/10/2013 07:15:00 [1] 109.0
13/10/2013 07:20:00 [1] 109.3
13/10/2013 07:25:00 [1] 109.6
13/10/2013 07:30:00 [1] 109.9
13/10/2013 07:35:00 [1] 110.2
13/10/2013 07:40:00 [1] 110.5
13/10/2013 07:45:00 [1] 110.8
13/10/2013 07:50:00 [1] 111.1
13/10/2013 07:55:00 [1] 111.4
13/10/2013 08:00:00 [1] 111.7
13/10/2013 08:05:00 [1] 112.0
13/10/2013 08:10:00 [1] 112.3
13/10/2013 08:15:00 [1] 112.6
13/10/2013 08:20:00 [1] 112.9
13/10/2013 08:25:00 [1] 113.2
13/10/2013 08:30:00 [1] 113.5
13/10/2013 08:35:00 [1] 113.8
13/10/2013 08:40:00 [1] 114.1
13/10/2013 08:45:00 [1] 114.4
13/10/2013 08:50:00 [1] 114.7
13/10/2013 08:55:00 [1] 115.0
13/10/2013 09:00:00 [1] 115.3
13/10/2013 09:05:00 [1] 115.6
13/10/2013 09:10:00 [1] 115.9
13/10/2013 09:15:00 [1] 116.2
13/10/2013 09:20:00 [1] 116.5
13/10/2013 09:25:00 [1] 116.8
13/10/2013 09:30:00 [1] 117.1
13/10/2013 09:35:00 [1] 117.4
13/10/2013 09:40:00 [1] 117.7
13/10/2013 09:45:00 [1] 118.0
13/10/2013 09:50:00 [1] 118.3
13/10/2013 09:55:00 [1] 118.6
13/10/2013 10:00:00 [1] 118.9
13/10/2013 10:05:00 [1] 119.2
13/10/2013 10:10:00 [1] 119.5
13/10/2013 10:15:00 [1] 119.8
13/10/2013 10:20:00 [1] 120.1
13/10/2013 10:25:00 [1] 120.4
13/10/2013 10:30:00 [1] 120.7
13/10/2013 10:35:00 [1] 121.0
13/10/2013 10:40:00 [1] 121.3
13/10/2013 10:45:00 [1] 121.6
13/10/2013 10:50:00 [1] 121.9
13/10/2013 10:55:00 [1] 122.2
13/10/2013 11:00:00 [1] 122.5
13/10/2013 11:05:00 [1] 122.8
13/10/2013 11:10:00 [1] 123.1
13/10/2013 11:15:00 [1] 123.4
13/10/2013 11:20:00 [1] 123.7
13/10/2013 11:25:00 [1] 124.0
13/10/2013 11:30:00 [1] 124.3
13/10/2013 11:35:00 [1] 124.6
13/10/2013 11:40:00 [1] 124.9
13/10/2013 11:45:00 [1] 125.2
13/10/2013 11:50:00 [1] 125.5
13/10/2013 11:55:00 [1] 125.8
13/10/2013 12:00:00 [1] 126.1
13/10/2013 12:05:00 [1] 126.4
13/10/2013 12:10:00 [1] 126.7
13/10/2013 12:15:00 [1] 127.0
13/10/2013 12:20:00 [1] 127.3
13/10/2013 12:25:00 [1] 127.6
13/10/2013 12:30:00 [1] 127.9
13/10/2013 12:35:00 [1] 128.2
13/10/2013 12:40:00 [1] 128.5
13/10/2013 12:45:00 [1] 128.8
13/10/2013 12:50:00 [1] 129.1
13/10/2013 12:55:00 [1] 129.4
13/10/2013 13:00:00 [1] 129.7
13/10/2013 13:05:00 [1] 130.0
13/10/2013 13:10:00 [1] 130.3
13/10/2013 13:15:00 [1] 130.6
13/10/2013 13:20:00 [1] 130.9
13/10/2013 13:25:00 [1] 131.2
13/10/2013 13:30:00 [1] 131.5
13/10/2013 13:35:00 [1] 131.8
13/10/2013 13:40:00 [1] 132.1
13/10/2013 13:45:00 [1] 132.4
13/10/2013 13:50:00 [1] 132.7
13/10/2013 13:55:00 [1] 133.0
13/10/2013 14:00:00 [1] 133.3
13/10/2013 14:05:00 [1] 133.6
13/10/2013 14:10:00 [1] 133.9
13/10/2013 14:15:00 [1] 134.2
13/10/2013 14:20:00 [1] 134.5
13/10/2013 14:25:00 [1] 134.8
13/10/2013 14:30:00 [1] 135.1
13/10/2013 14:35:00 [1] 135.4
13/10/2013 14:40:00 [1] 135.7
13/10/2013 14:45:00 [1] 136.0
13/10/2013 14:50:00 [1] 136.3
13/10/2013 14:55:00 [1] 136.6
13/10/2013 15:00:00 [1] 136.9
13/10/2013 15:05:00 [1] 137.2
13/10/2013 15:10:00 [1] 137.5
13/10/2013 15:15:00 [1] 137.8
13/10/2013 15:20:00 [1] 138.1
13/10/2013 15:25:00 [1] 138.4
13/10/2013 15:30:00 [1] 138.7
13/10/2013 15:35:00 [1] 139.0
13/10/2013 15:40:00 [1] 139.3
13/10/2013 15:45:00 [1] 139.6
13/10/2013 15:50:00 [1] 139.9
13/10/2013 15:55:00 [1] 140.2
13/10/2013 16:00:00 [1] 140.5
13/10/2013 16:05:00 [1] 140.8
13/10/2013 16:10:00 [1] 141.1
13/10/2013 16:15:00 [1] 141.4
13/10/2013 16:20:00 [1] 141.7
13/10/2013 16:25:00 [1] 142.0
13/10/2013 16:30:00 [1] 142.3
13/10/2013 16:35:00 [1] 142.6
13/10/2013 16:40:00 [1] 142.9
13/10/2013 16:45:00 [1] 143.2
13/10/2013 16:50:00 [1] 143.5
13/10/2013 16:55:00 [1] 143.8
13/10/2013 17:00:00 [1] 144.1
13/10/2013 17:05:00 [1] 144.4
13/10/2013 17:10:00 [1] 144.7
13/10/2013 17:15:00 [1] 145.0
13/10/2013 17:20:00 [1] 145.3
13/10/2013 17:25:00 [1] 145.6
13/10/2013 17:30:00 [1] 145.9
13/10/2013 17:35:00 [1] 146.2
13/10/2013 17:40:00 [1] 146.5
13/10/2013 17:45:00 [1] 146.8
13/10/2013 17:50:00 [1] 147.1
13/10/2013 17:55:00 [1] 147.4
13/10/2013 18:00:00 [1] 147.7
13/10/2013 18:05:00 [1] 148.0
13/10/2013 18:10:00 [1] 148.3
13/10/2013 18:15:00 [1] 148.6
13/10/2013 18:20:00 [1] 148.9
13/10/2013 18:25:00 [1] 149.2
13/10/2013 18:30:00 [1] 149.5
13/10/2013 18:35:00 [1] 149.8
13/10/2013 18:40:00 [1] 150.1
13/10/2013 18:45:00 [1] 150.4
13/10/2013 18:50:00 [1] 150.7
13/10/2013 18:55:00 [1] 151.0
13/10/2013 19:00:00 [1] 151.3
13/10/2013 19:05:00 [1] 151.6
13/10/2013 19:10:00 [1] 151.9
13/10/2013 19:15:00 [1] 152.2
13/10/2013 19:20:00 [1] 152.5
13/10/2013 19:25:00 [1] 152.8
13/10/2013 19:30:00 [1] 153.1
13/10/2013 19:35:00 [1] 153.4
13/10/2013 19:40:00 [1] 153.7
13/10/2013 19:45:00 [1] 154.0
13/10/2013 19:50:00 [1] 154.3
13/10/2013 19:55:00 [1] 154.6
13/10/2013 20:00:00 [1] 154.9
13/10/2013 20:05:00 [1] 155.2
13/10/2013 20:10:00 [1] 155.5
13/
```

# Temporal data mining : PEPS project

Développement d'une plate-forme pour mener des études de pharmaco-épidémiologies à partir des données de l'Assurance Maladie



# Temporal data mining : PEPS project

Développement d'une plate-forme pour mener des études de pharmaco-épidémiologies à partir des données de l'Assurance Maladie

Accélérer la phase initiale d'exploration des données Besoins d'outils pour manipuler facilement les parcours de soins

- Permettre l'expression de requêtes complexes
- Intégrer des outils de requête et de fouille
- Lier les outils aux méthodes de visualisation

Aider à la formulation de nouvelles hypothèses

- Besoin de dépasser le test statistique de co-occurrences
- Caractéristiques des parcours de soins : identification de séquences de consommations et d'effets
- Associations entre parcours et caractéristiques de patient

# Focusing research on data mining

Les questions de **fouille de données** deviennent centrales dans nos activités, mais toujours dans une approche de l'**Intelligence artificielle**.

Vers une nouvelle équipe de recherche portée par **A. Termier**.

## Thèmes d'intérêt (*non-contractuel*;) )

- Automating the exploration of the KDD search space
  - Collaborative knowledge and feedback management
  - Scaling up through in-memory approaches
  - User/system interactions
- ⇒ Large Collaborative Data Mining

# Outline

- 1 Monitoring dynamical systems and data mining in Rennes
  - DREAM Team
  - Monitoring dynamical systems
  - Quelques applications
  - Réorganisation des objectifs de recherche
- 2 Une approche exemple de Pattern Mining et IA : extraction de motifs séquentiels avec ASP
  - Pattern mining
  - Fouille avec ASP
  - Cas exemple : analyse de trace de patients
- 3 Pattern mining et Intelligence Artificielle : Questions ouvertes
  - Planification et extraction de motifs
  - Composer des chaînes d'extraction de connaissances
  - Raisonnement (décision/argumentation) sur et avec les motifs



# Présentation d'une approche exemple

On présente ici une approche déclarative de l'extraction de motifs

- rapprochement vers la résolution de tâches formelles (telles que la planification ou autre?)
- premier pas de rapprochement vers l'IA : utilisation d'outils de raisonnement automatique (*answer set programming*) pour résoudre la tâche
- perspective d'intégration de connaissances et d'heuristiques dans la résolution

# Frequent pattern mining

- A sub-task of the data mining field
- The task
  - Patterns : itemsets (a set of items), sequences, graphs
  - Data : database of structured transactions, e.g. itemsets, sequences
  - Goal : **extracting all the frequent patterns in a database.**
- Frequent? The number of transaction that support the pattern is above a given threshold  $\sigma$
- Classical algorithms
  - Itemsets : Apriori, FP-Growth, LCM, ...
  - Sequential patterns : GSP, PrefixSpan, SPADE, BiDE, ...
  - Graphs : gSpan

## Frequent patterns : itemsets

TID	itemset
10	a, b, c
20	a, c, d
30	a, d
40	b, e, f

- frequency threshold : 2
- set of frequent patterns :  
 $\{a, b, c, d, ac, ad\}$

# Frequent pattern mining

- A sub-task of the data mining field
- The task
  - Patterns : itemsets (a set of items), sequences, graphs
  - Data : database of structured transactions, e.g. itemsets, sequences
  - Goal : **extracting all the frequent patterns in a database.**
- Frequent? The number of transaction that support the pattern is above a given threshold  $\sigma$
- Classical algorithms
  - Itemsets : Apriori, FP-Growth, LCM, ...
  - Sequential patterns : GSP, PrefixSpan, SPADE, BiDE, ...
  - Graphs : gSpan

## Frequent patterns : sequences

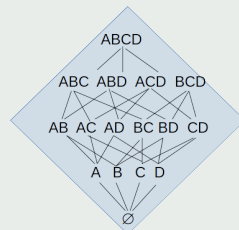
SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)abc>

- frequency threshold : 2
- set of frequent patterns :  
 $\{a, ab, ac, ad, af, (ab), abc, \dots\}$

# Mining frequent patterns : algorithms principles I

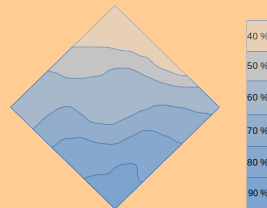
## Search space of patterns

- patterns are ordered according to a partial order (inclusion relation,  $\subset$ )
- ⇒ lattice of patterns



## Anti-monotony property of the frequency

- Let  $P$  be a pattern and  $supp(P)$  its support.  $\forall Q \subset P$ , we have  $supp(Q) \geq supp(P)$ .
- ⇒ the frequency of a pattern decrease with its depth in the lattice



IRISA

# Mining frequent patterns : algorithms principles II

## Frequent pattern mining

Mining all the pattern whose support is above a given threshold  $\sigma$ .

- support : number of transaction in the database that support the pattern
- a transaction supports a pattern if it "contains" the pattern
  - itemsets : subset of a transaction
  - sequences : sub-sequence of a transaction (sequence) with gaps !

# Mining frequent patterns : algorithms principles III

## Two most important issues of the pattern mining algorithms

### 1 search space traversal

- ⇒ use the anti-monotony property to cut the search as soon as possible
- ⇒ the browsing must be efficient and it avoids redundant evaluation of pattern

### 2 support evaluation

- ⇒ reduce the number of access to the database
- ⇒ split the databases in sub-databases (enable efficient parallelism)
- ⇒ pattern-transaction matching algorithm (simple for itemsets, but may be harder, e.g. graphs)

# Pattern mining and Declarative Programming

## "Main" principles of Declarative Programming

- 1 a set of "generators" that generates potential answers ← **traversal of the search space**
- 2 a set of "constraints" that defines what is a "valid" model and prunes the others ← **support evaluation**
- 3 optimization criteria ← **pattern selection**

- Constraint programming approaches []
- Answer Set Programming [] : appropriate to integrate knowledge

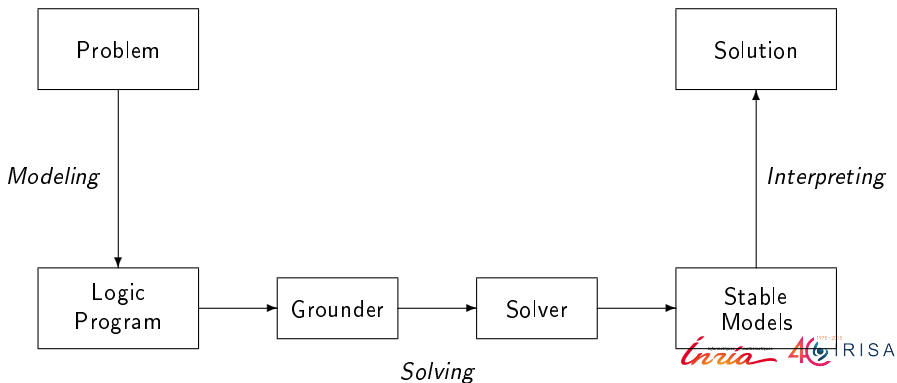
## Our objectives

- finding efficient declarative encoding of the pattern mining tasks
  - **integrating reasoning** into the mining process
- ⇒ especially applied on the **sequential pattern mining** issue

# Answer Set Programming

## Programme ASP

- idéal de la programmation déclarative : *le problème est le programme*
- la solution est un ensemble de modèles (ensembles réponses)
- utilisation de contraintes sur des ensembles





# Programme ASP

## Principe "général" de programmation

- ① ensemble de règles pour "générer" des modèles (réponses potentielles)
  - ② ensemble de contraintes : éliminent les modèles non-souhaités
  - ③ directive d'optimisation : `#minimize`, `#maximize`
  - ④ directive d'affichage : `#show`
- ⇒ syntaxe proche du prolog (symboles de premier ordre) faciles à écrire

## Coloration de graphe

```
node(1). node(2). node(3). node(4). node(5). node(6).
edge(1,2). edge(1,3). edge(1,4). edge(2,4). edge(2,5). edge(2,6).
edge(3,1). edge(3,4). edge(3,5). edge(4,1). edge(4,2). edge(5,3).
edge(5,4). edge(5,6). edge(6,2). edge(6,3). edge(6,5).
col(r). col(b). col(g).
```

```
1 { color(X,C) : col(C) } 1 :- node(X).
:- edge(X,Y), color(X,C), color(Y,C).
```

```
#show color/2.
```

# Fouille d'itemsets avec ASP

## Programme ASP proposé par Järvisalo [Jär11]

```

item(I) :- db(_,I). % ensemble des items de la base de transactions
transaction(T) :- db(T,_). % ensemble des transactions de la base

{ in_itemset(I) } :- item(I). % generation d'un motif par AS
in_support(T) :- { conflict_at(T,I) : item(I) } 0, transaction(T).
conflict_at(T,I) :- not db(T,I), in_itemset(I), transaction(T).
:- { in_support(T) } N-1, threshold(N).
  
```

- `item/1` : liste d'items
- `transaction/1` : liste des transaction
- `db/2` : contenu de la matrice

	Item(1)	Item(2)	Item(3)	Item(4)	Item(5)
Transaction(1)	■	■	■	■	■
Transaction(2)	■	■	■	■	■
Transaction(3)	■	■	■	■	■
Transaction(4)	■	■	■	■	■
Transaction(5)	■	■	■	■	■
Transaction(6)	■	■	■	■	■
Transaction(7)	■	■	■	■	■
Transaction(8)	■	■	■	■	■

# Fouille d'itemsets avec ASP

Programme ASP proposé par Järvisalo [Jär11]

```
item(I) :- db(_,I). % ensemble des items de la base de transactions
transaction(T) :- db(T,_). % ensemble des transactions de la base

{ in_itemset(I) } :- item(I). % generation d'un motif par AS
in_support(T) :- { conflict_at(T,I) : item(I) } 0, transaction(T).
conflict_at(T,I) :- not db(T,I), in_itemset(I), transaction(T).
:- { in_support(T) } N-1, threshold(N).
```

- il propose comme principe : 1 AS = 1 motif
- il reprend la modélisation basée sur l'utilisation de conflit de [GNR11] : `conflict_at/2` indique qu'un motif n'est pas supporté par une transaction `T`

# Fouille d'itemsets avec ASP

## Programme ASP proposé par Järvisalo [Jär11]

```
item(I) :- db(_,I). % ensemble des items de la base de transactions
transaction(T) :- db(T,_). % ensemble des transactions de la base

{ in_itemset(I) } :- item(I). % generation d'un motif par AS
in_support(T) :- { conflict_at(T,I) : item(I) } 0, transaction(T).
conflict_at(T,I) :- not db(T,I), in_itemset(I), transaction(T).
:- { in_support(T) } N-1, threshold(N).
```

## Quid de la propriété d'anti-monotonie ?

- Le solveur "apprend" la relation d'anti-monotonie
  - Expérimentations d'alternatives peu concluantes (approche incrémentale, modification des heuristiques de recherche)
- ⇒ déterminer une bonne heuristique de parcours de l'espace de recherche des motifs

# ASP : un outil pour mixer fouille de données et intelligence artificielle

- le programme précédent résoud la tâche de fouille d'itemsets (relativement efficacement)
- il est très aisé d'ajouter des connaissances sur les motifs souhaités
  - contraintes sur les items : présence d'un item, obliger/interdire deux items à être ensemble,
  - contraintes sur l'ensemble de motif : difficile avec cette solution.
- connaissances modélisant des informations complémentaire sur les items (e.g. le prix de chaque item)
  - ⇒ permet de définir des contraintes sémantiquement très riches
- il est possible de manipuler l'heuristique de résolution
  - ⇒ sortir les motifs "les plus intéressants" en premier

```
cost(1,12).cost(2,23).cost(3,56). ...  
:- L=#sum{C : cost(I,C), in_itemset(I)}, L>100.
```

# Motifs séquentiels

## Extraction de motifs séquentiels

La fouille de séquences est plus intéressantes algorithmiquement parlant et facilite l'expression d'information riches !

- défis de modélisations ASP ouvrant vers la résolution de tâches difficiles de fouille
- démonstratif de l'approche déclarative

## Encodage de la base de séquences

- $\text{seq}(T,D,I)$  : transaction  $T$ , at date  $D$ , the item  $I$

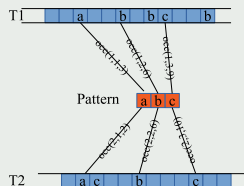
```
seq(0,1,20) . seq(0,2,19) . seq(0,3,2) . seq(0,4,18) .  
seq(1,1,5) . seq(1,2,6) . seq(1,3,6) . seq(1,4,14) .  
seq(2,1,13) . seq(2,2,12) . seq(2,3,9) . seq(2,4,19) .  
seq(3,1,15) . seq(3,2,7) . seq(3,3,17) . seq(3,4,17) .  
seq(4,1,11) . seq(4,2,2) . seq(4,3,2) . seq(4,4,13) .  
seq(5,1,14) . seq(5,2,8) . seq(5,3,13) . seq(5,4,1) .
```

# Motifs séquentiels I

Version d'un encodage inspirée de l'approche Järvisalo.

- `patlen/1` donne la longueur des motifs (fixé ici)
- `pat/2` décrit le motif (1 motif/AS)
- `occ/3` décrit les occurrences du motif
- `covered_transaction/1` décrit les transactions couvertes par le motif

Exemple : description des occurrences d'une transaction



```
pat(a). pat(b). pat(c).
occ(1,1,3). occ(1,2,6). occ(1,3,9).
occ(2,1,2). occ(2,2,6). occ(2,3,10).
```

# Motifs séquentiels II

```

1 transaction (T) :- seq(T,_,_).
2 symb(S) :- seq(_,_,S).
3 1 { patlen (1.. maxlen) } 1.
4 pat (1.. L, C): symb(C) :- patlen(L).
5
6 th { covered_transaction(T) : transaction (T) } th.
7
8 % I / T is the instance id, sequence id
9 % L est la longueur du pattern, P position in the sequence,
10 % C: the item
11 occ(T, 1.. L, P): seq(T,P,C) :- patlen(L), covered_transaction(T).
12 :- occ(I,N,P), occ(I,N,Q), P<Q.
13 :- occ(I,N,P), pat(N,C), not seq(I,P,C) .
14
15 % respect de l'ordre dans les patterns :
16 :- occ(I, N, P), occ(I, N2, P2), N<N2, P>=P2.

```



# Expérimentation en cours I

## Cas d'application : fouille de parcours de soins

- Base de données de l'assurance maladie
  - information détaillée sur les remboursements de l'ensemble des assurés sociaux
    - médicaments (date de délivrance, quantité, etc.)
    - passage à l'hôpital (identification de pathologies)
  - les remboursements sont codifiés (taxonomie des médicaments)
- Données de l'étude GenEpi : étude des parcours de soin de patient épileptiques avant une hospitalisation pour crise convulsive
- Objectif : identification de parcours "expliquant" des crises

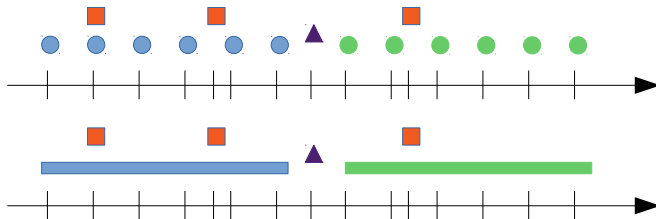
## Données

- Nombres de codes CIPs : 1208
- Nombre de transaction : 100
- nb items/transaction : 243.16
- nb items différents/transaction : 48.24

# Expérimentation en cours II

## Les défis

- modélisation d'une taxonomie riche
  - inclusion d'étapes de pré-traitements : regroupement des délivrances régulières de médicaments
- ⇒ approches encore en cours de développement



# Outline

- 1 Monitoring dynamical systems and data mining in Rennes
  - DREAM Team
  - Monitoring dynamical systems
  - Quelques applications
  - Réorganisation des objectifs de recherche
- 2 Une approche exemple de Pattern Mining et IA : extraction de motifs séquentiels avec ASP
  - Pattern mining
  - Fouille avec ASP
  - Cas exemple : analyse de trace de patients
- 3 **Pattern mining et Intelligence Artificielle : Questions ouvertes**
  - Planification et extraction de motifs
  - Composer des chaînes d'extraction de connaissances
  - Raisonnement (décision/argumentation) sur et avec les motifs

# Espaces de recherche en extraction de motifs

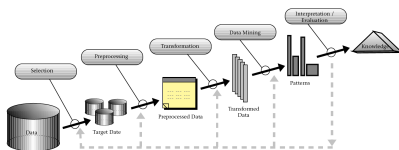
- le problème d'extraction de motifs est-il semblable à un problème de planification ?
  
- l'intérêt de répondre positivement à cette question serait de bénéficier des travaux en planification pour appliquer des heuristiques (complètes ou non) à l'extraction de motifs
  - ⇒ planification sous contraintes = motifs contraints ?

# Espaces de recherche en extraction de motifs

- le problème d'extraction de motifs est-il semblable à un problème de planification ?
  - Un motif est-il un "plan" ? (séquence  $\rightarrow$  planification séquentielle ?)
  - Un motif est-il une "allocation de ressource" ?
  - Les contraintes sur les "plans" sont liées à la base de données ! $\Rightarrow$  Comment considérer la contrainte forte de limitation d'un accès aux données
- l'intérêt de répondre positivement à cette question serait de bénéficier des travaux en planification pour appliquer des heuristiques (complètes ou non) à l'extraction de motifs
  - $\Rightarrow$  planification sous contraintes = motifs contraints ?

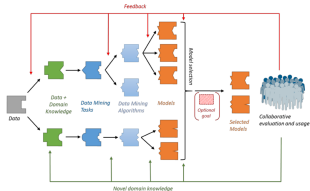
# Composer des chaînes d'extraction de connaissances

- les outils de patterns mining ne sont qu'une étape du processus de fouille de données
- beaucoup d'autres méthodes doivent être mise en oeuvre pour extraire de l'information des données brutes
- le schéma de Fayyad est très linéaire et ne montre pas la difficulté de l'utilisateur à concevoir et interagir avec cette chaîne de traitements



# Composer des chaînes d'extraction de connaissances

- les outils de patterns mining ne sont qu'une étape du processus de fouille de données
- beaucoup d'autres méthodes doivent être mise en oeuvre pour extraire de l'information des données brutes
- le schéma de Fayyad est très linéaire et ne montre pas la difficulté de l'utilisateur à concevoir et interagir avec cette chaîne de traitements
- l'ensemble des outils d'analyse possibles et leurs paramétrisation fait de la question de la conception d'une chaîne de traitement un problème hautement combinatoire



# Composer des chaînes d'extraction de connaissances

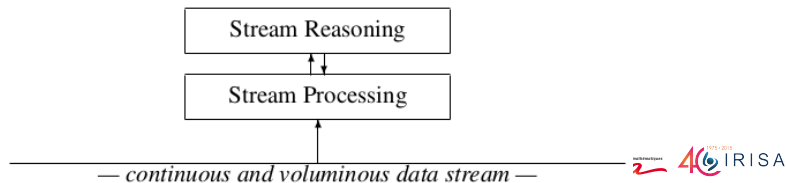
- les outils de patterns mining ne sont qu'une étape du processus de fouille de données
- beaucoup d'autres méthodes doivent être mise en oeuvre pour extraire de l'information des données brutes
- le schéma de Fayyad est très linéaire et ne montre pas la difficulté de l'utilisateur à concevoir et interagir avec cette chaîne de traitements
- l'ensemble des outils d'analyse possibles et leurs paramétrisation fait de la question de la conception d'une chaîne de traitement un problème hautement combinatoire
- l'utilisation de méthodes de planification permettrait d'automatiser en partie cette étape
  - Automatisation de la construction d'un plan d'analyse de données guidée par un "but d'analyse"
  - Suggestions de paramétrisation
  - Apprentissage à partir de plans
  - Généralisation de plans d'analyse de données



# Mieux choisir les motifs I

## Constat

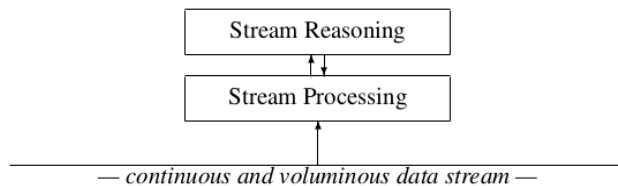
- les méthodes d'extraction de motifs génèrent un très grand nombre de motifs
  - besoin d'affiner l'ensemble des motifs présentés à l'utilisateur
    - a priori : trouver des méthodes capables d'intégrer des connaissances utilisables pendant la fouille
    - a posteriori : filtrage de l'ensemble des motifs intéressants (toujours en utilisant des connaissances du domaine)
      - ⇒ bénéficier de l'efficacité des outils de fouille
      - ⇒ utiliser de la connaissance du domaine pour filtrer les motifs
- ⇒ mise en œuvre de raisonnements sur les motifs



# Mieux choisir les motifs II

## Prendre des décisions sur les motifs

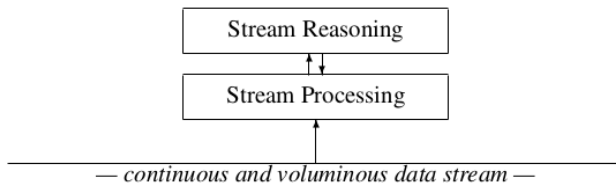
- Sachant un ensemble de motifs et des connaissances expertes,
    - quelles sont les motifs ou l'ensemble de motifs les plus "intéressants"  
exemple de critères : actionnabilité des actions, ...
    - comment les structurer (préférences)
- ⇒ comment comparer des motifs ?
- ⇒ vers des réflexions sur l'utilisation de méthodes d'argumentation : pourquoi un motif est-il plus intéressant ?



# Mieux choisir les motifs III

## Prendre des décisions *avec* des motifs

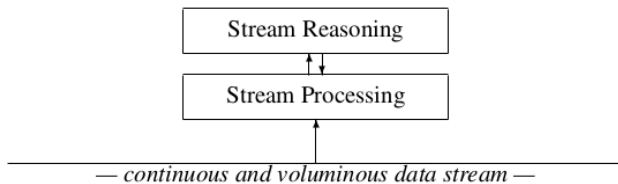
- comment un ensemble de motifs peut être utilisé pour prendre des décisions en ligne?
- problème : quelle est la pertinence des *motifs fréquents* pour prendre des décisions ?



# Mieux choisir les motifs IV

## Prise de décision d'ordre 2

Décider de quand et sur quoi renouveler l'ensemble de motifs (en cas *concept drift*)



# Conclusion

- ⇒ L'extraction de motifs dans des bases de données ouvrent des questions intéressantes pour des méthodes d'intelligence artificielle
- ⇒ Nos activités nous conduisent à traiter de plus en plus de données ... un des défis majeurs sera faire passer les méthodes d'IA à l'échelle de ces données !

Questions ?

# References |



Tias Guns, Siegfried Nijssen, and Luc De Raedt, *Itemset mining : A constraint programming perspective*, Artificial Intelligence 175 (2011), no. 12-13, 1951–1983.



Matti Järvisalo, *Itemset mining as a challenge application for answer set enumeration*, Logic Programming and Nonmonotonic Reasoning, Springer, 2011, pp. 304–310.



Simon Malnowski, Thomas Guyet, and René Quiniou, *Proceedings of the Intelligence Data Analysis*, 2013.