

# On-Line Feature Selection in Data Streams: Applications to Clustering and Usage Mining

**Keywords :** Data Mining and Knowledge Discovery; Data Streams; Information Theory; Feature Selection; Web Usage Mining; Change Detection.

**Team:** The AxIS project-team (INRIA) aims at improving information systems thanks to usage analysis techniques (such as Web usage mining for instance). AxIS already proposed numerous solutions for stream mining (frequent pattern extraction, clustering, summarizing) and feature selection. AxIS activities also focus on recommender systems and case based reasoning. A main characteristic of AxIS is to provide solutions coming from multiple domains such as statistics, data mining, machine learning, ergonomics or software engineering.

## Subject:

The subject involves the following domains :

1. **Dimensionality reduction.** This is an important problem in domains related to information processing and knowledge extraction such as summarizing data, searching and indexing Web data, classifying and clustering, etc. Feature Selection and Feature Extraction are the main approaches of dimensionality reduction. Feature Selection techniques are designed to tackle the high computational problem associated to large-scale or streaming data. Their goal is to find out a subset of features that are as representative as possible according to their predictive power.
2. **Data Streams.** In this domain, data arrive in a potentially infinite stream, at a very high rate and it is not allowed to perform blocking operations. It is not possible to record the data of a stream and associated methods have to face many challenges.

**Feature Selection in Data Streams** generally aims at maintaining an optimal subset of features on-line. The goal is still to reduce the CPU time needed for processing the stream but also to detect a possible change in the existing model. Actually some less relevant features might become representative with the passing of time and the associated algorithms (classification, prediction, etc.) should lose effectiveness and/or efficiency if this is not detected. Another application of on-line feature selection in data streams is detecting concept drift (the concept of the distribution in a class may change).

**Supervisor:** [Florent Masseglia](mailto:Florent.Masseglia@sophia.inria.fr) (Florent.Masseglia@sophia.inria.fr)

## Skills and profile:

- Previous knowledge about data mining and/or information theory.
- Good skills in C++ and/or Java.
- Good proficiency in English
- M.Sc in Computer Science.

The student will work on the French Riviera, in the [INRIA center of Sophia-Antipolis](#).

Net salary: 1537 €/month the first two years and 1619 €/month the 3<sup>rd</sup> year.

Send a detailed CV (including results of the M.Sc), a letter presenting your motivation for this Ph.D and at least two reference letters before April 30.