# MDL for Robust Periodic Pattern Mining
# Or How to Detect Habit Changes?

Peggy Cellier, IRISA - SemLIS, INSA Rennes, peggy.cellier@irisa.fr
Esther Galbrun, University of Eastern Finland, esther.galbrun@uef.fi
Alexandre Termier, IRISA - Lacodam, Univ. de Rennes 1 Alexandre.Termier@irisa.fr

**Location**: IRISA, team Lacodam/SemLIS, Rennes

**Keywords**: data science, data mining, pattern mining, Minimum Description Length (MDL)

**Context and description of the internship**:

Event logs are among the most ubiquitous types of data nowadays. They can be machine generated (e.g., sensor data) or human generated (ranging from hospital records to life tracking, a.k.a. quantified self), and are bound to become ever more voluminous and diverse with the increasing digitisation of our lives and the advent of the Internet of Things (IoT). Such logs are often the most readily available sources of information about a system or process of interest. It is thus critical to have effective and efficient means to analyse them and extract the information they contain.

Uncovering hidden patterns in those logs is an important analysis task and the goal of the *pattern mining* research field. The information captured can for example be co-occurrences (e.g. *'wake up' and 'prepare coffee' events often appear together in the log*) or order relations (e.g. *'wake up' events are often* ***followed by*** *'prepare coffee' events in the log*). In logs, this is often not sufficient, and the time component has to be taken into account more specifically. This led to the proposition of *periodic patterns* [1,2], which allow to discover patterns appearing with some temporal regularity (e.g. ***(Nearly) every 24h****, there are 'wake up' events* ***followed by*** *'prepare coffee' events in the log*). In a recent work [3], we proposed a first method for discovering *nested* periodic patterns, which allows to discover complex repetitions of the form:

> « ***Starting Monday at 7:30 AM****,*
> *wake up*, ***then, 10 minutes later,*** *prepare coffee,*
> ***repeat every 24 hours for 5 days,***
> ***repeat this every 7 days for 3 months*** »

Such kind of patterns, when applied over logs of a person's activities, can be used to detect their *habits*, while over system logs they can help characterize the standard regimes of operation of the system.

Now, an interesting question is: *can we detect slow and gradual changes of behaviors of the person or system with these patterns?*

Such changes are called **concept drift** and can be particularly difficult to detect, especially when the changes are very gradual. The approach we envision would offer the following advantages:1) the habit is not specified beforehand and is automatically detected by pattern mining; and 2) the change can be precisely qualified by reporting to the analyst which part of the pattern is being gradually modified.

The goal of the internship is to revisit the approach proposed in [3] in order to detect nested periodic pattern that exhibit concept drift in historical data. An important challenge will be to output only a few meaningful patterns. In [3], we rely on an approach from Information Theory, the Minimum Description Length (MDL) principle [8], which has been successfully used in pattern mining approaches [4-7]. While it has been proven that MDL combined with pattern mining could be used to detect simple cases of abrupt changes in streaming data [9], it has never been used to detect the more subtle cases of concept drift aimed at in this internship.

The expected research work requires a taste for theory, algorithm design and experiments. This internship subject can lead to a PhD.

[1] B. Özden, S. Ramaswamy, and A. Silberschats. **Cyclic association rules**. ICDE, 1998.

[2] P. Lopez-Cueva, A. Bertaux, A. Termier, J .-F. Méhaut, and M. Santana. **Debugging embedded multimedia application traces through periodic pattern mining**. Int. Conf. On Embedded Software, 2012.

[3] Esther Galbrun, Peggy Cellier, Nikolaj Tatti, Alexandre Termier, and Bruno Crémilleux. **Mining Periodic Patterns with a MDL Criterion**. ECML-PKDD 2018.

[4] J. Vreeken, M. van Leeuwen, and A. Siebes. **Krimp : Mining itemsets that compress**. DMKD, 2011.

[5] F. Bonchi, M. van Leeuwen, and A. Ukkonen. **Characterizing uncertain data using compression**. SDM, 2011.

[6] N. Tatti and J. Vreeken. **The long and the short of it: Summarising event sequences with serial episodes**. KDD, 2012.

[7] A. Bhattacharyya and J. Vreeken. **Efficiently summarising event sequences with rich interleaving patterns**. SDM 2017.

[8] P. Grünwald. **Model Selection Based on Minimum Description Length**. Journal of Mathematical Psychology, 2000.

[9] M. van Leeuwen and A. Siebes. **StreamKrimp: Detecting Change in Data Streams**. ECML/PKDD (1) 2008: 672-687