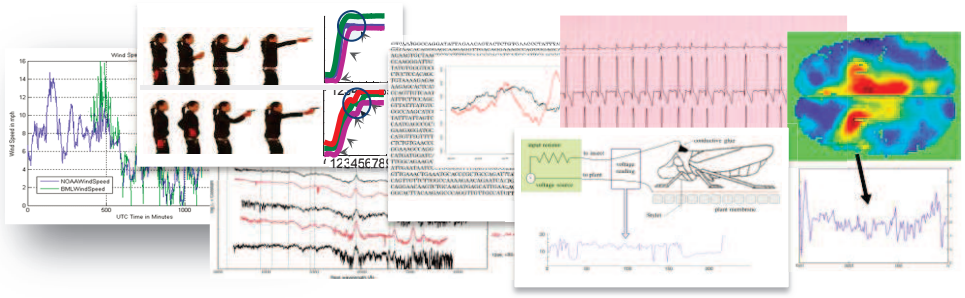
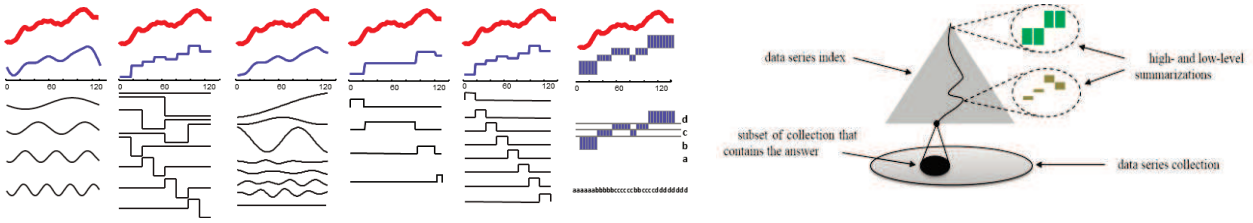
# Parallelization for Ultra-Fast Data Series Indexing

***Advisor: Prof. Themis Palpanas (Paris Descartes University)***
*themis@mi.parisdescartes.fr*

The development of sensor technologies in a wide range of domains (e.g., earth observation, astronomy, genome sequencing) has led to an explosion in monitoring activities, which provide a very large amount of data series (i.e., ordered sequences of values).



In order to efficiently process and analyze large volumes of data series, we have to operate on summaries (or approximations) of these data series. Several techniques have been proposed in the literature for the approximation of data series [1], including Discrete Fourier Transform (DFT), Piecewise Aggregate Approximation (PAA), Discrete Wavelet Transform (DWT), Symbolic Aggregate approXimation (SAX), and others.




Based on these approximations, we can then build indexes that help us answer fast similarity queries on massive collections of data series. Our group has developed the current state of the art data series indexes [2][3]: we have been able to experimentally demonstrate scalability to dataset sizes of 1 billion data series, which is 2-3 orders of magnitude more than the previous approaches. Nevertheless, new revolutionary techniques that also facilitate new hardware capabilities are necessary in order to further improve scalability.

The goal of this project is to optimize the performance of these indexes by using multi-cores and modern CPU capabilities such as SIMD instructions. Such enhancements can be applied in both the index building step, as well as the query answering step. Processing data series requires a large number of numerical computations, but at the same time is a problem amenable to parallelization. As a result, we expect that CPU optimizations will bring significant benefits to the overall performance.

Accepting this project will make you part of an enthusiastic team working on real, challenging problems!
Prerequisites: experience with file and data structures, excellent programming skills.

**References**

[1] Themis Palpanas, Michail Vlachos, Eamonn Keogh, Dimitrios Gunopulos. Streaming Time Series Summarization Using User-Defined Amnesic Functions. TKDE 20(7), 2008.
[2] Alessandro Camerra, Jin Shieh, Themis Palpanas, Thanawin Rakthanmanon, Eamonn Keogh. Beyond One Billion Time Series: Indexing and Mining Very Large Time Series Collections with iSAX2+. KAIS 39(1), 2014.
[3] Kostas Zoumpatianos, Stratos Idreos, Themis Palpanas. Indexing for Interactive Exploration of Big Data Series. SIGMOD 2014.