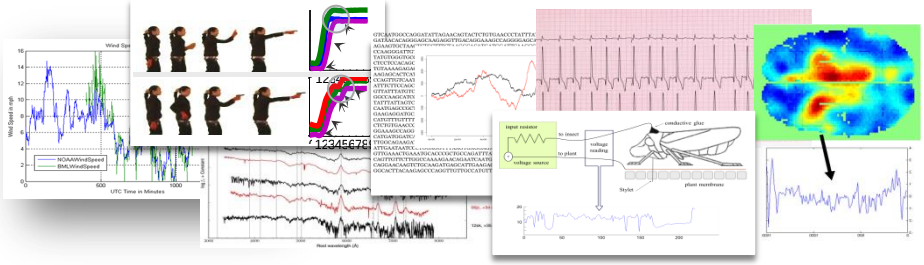
Prof. Themis Palpanas
Institut Universitaire de France
Paris Descartes University
themis@mi.parisdescartes.fr

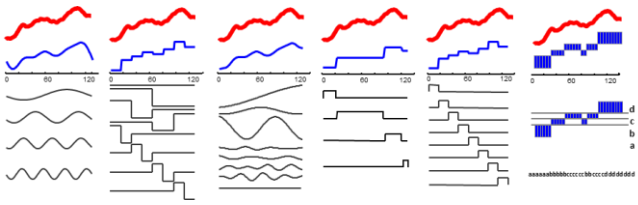
Funded Research M2 Internship (2nd year of MSc) for 2018:

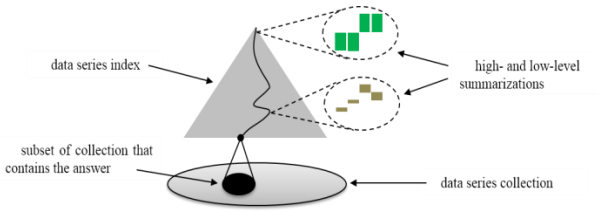
Machine Learning for Massive Data Series Collections

Context: There is an increasingly pressing need, by several applications in diverse domains, for developing techniques able to index and perform complex analytics on very large collections of data series (i.e., sequences of values) [1].



In order to efficiently process and analyze large volumes of data series, we have to operate on summaries (or approximations) of these data series, which are subsequently indexed in order to enable fast and scalable similarity search query answering. 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.





Problem: The purpose of this project is to design techniques for applying machine learning algorithms on truly massive collections of data series. This is particularly challenging, because several machine learning algorithms rely on distance computations and similarity search for their functionality, and it is exactly these operations that are extremely expensive to perform with data series objects, and especially when dealing with very large collections of data series. In this project, we will study the most prominent machine learning techniques, and propose novel, scalable algorithms that implement these techniques. We will also examine how data series indexes can be used to this effect, and to what extent they can contribute to the scalability of machine learning techniques. We will perform experimental evaluations with synthetic and real datasets, in order to measure the performance improvements.

Internship: Accepting this project will make you part of an enthusiastic team working on real, challenging problems!
Prerequisites: experience with machine learning techniques, file and data structures, excellent programming skills.
The internship will last 3-6 months, and is fully funded. We expect to publish a paper based on the obtained results.

Team:

Themis Palpanas is senior member of the Institut Universitaire de France, and professor of computer science at the Paris Descartes University, where he is co-director of diNo, the data management group. He received the MSc and PhD degrees from the University of Toronto, Canada. His team has developed world-wide expertise on data series management and analysis.

Kostas Zoumpatianos is a postdoctoral researcher at Paris Descartes university and Harvard University. He rceived his PhD from thh University of Trento. He has developed the current state of the art techniques for data series indexing.

References:

[1] Themis Palpanas: Big Sequence Management: A glimpse of the Past, the Present, and the Future. SOFSEM 2016: 63-80.
[2] Alessandro Camerra, Jin Shieh, Themis Palpanas, Thanawin Rakthanmanon, Eamonn J. Keogh: Beyond one billion time series: indexing and mining very large time series collections with iSAX2+. Knowl. Inf. Syst. 39(1): 123-151 (2014).
[3] Kostas Zoumpatianos, Stratos Idreos, Themis Palpanas: ADS: the adaptive data series index. VLDB J. 25(6): 843-866 (2016).