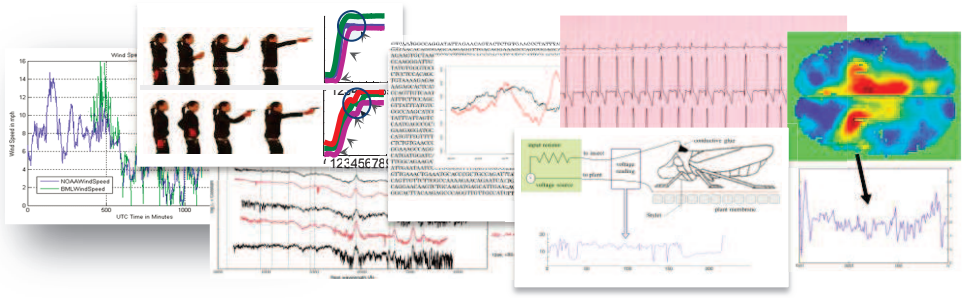
Advanced Analytics: Mining All Lag-Correlated Data Series

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. Such index structures are particularly useful in real-time monitoring for business intelligence [4][5].

The goal of this project is to monitor different measures in a data warehouse (e.g., sales of products over time), and identify time series that are similar/correlated, irrespective of the time-lag between these series. Such insights can provide the basis for algorithms that provide timely warnings. For example, based on the current pattern monitored in one time series, an algorithm could predict the future values of another time series. Previous research was able to produce rules in the form of simple increase, decrease operations: e.g., a decrease of 10% in housing prices leads to an increase of 10% in house sales. In this work, we will provide a much more detailed analysis, while being able to scale to a very large number of time series.

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.
[4] K. Zoumpatianos, T. Palpanas, J. Mylopoulos. Strategic Management for Real-Time Business Intelligence. BIRTE 2012.
[5] K. Zoumpatianos, T. Palpanas, J. Mylopoulos, Alejandro Mate, Juan Trujillo. Monitoring and Diagnosing Indicators for Business Analytics. CASCON 2013.