Time series data is naturally formed by repeated measurements over time and their analysis in various science and engineering disciplines precedes the digital age. With progressing digitalization the amount and detail of such data is ever more growing and drives the demand for performant data mining tools to gain insights from the increasing amounts of data. The matrix profile proposed is a data structure which can serve in a variety of time series data mining tasks like motif search or clustering. Such analysis can be performed with little effort given the matrix profile, but obtaining the latter is a computationally intensive task. In this work we investigate migration of the matrix profile computation to a high performance computing cluster utilizing the MPI standard. Highly optimized parallel routines of MPI and the vast number of available computing elements on such clusters allow users to scale their algorithms in terms of runtime and level of detail according to their needs within the most distant limits. We present an approach, which enables the analysis of previously intractably large time series based on the matrix profile. It provides nearly unlimited scalability to solve large problems within reasonable runtimes according to the users needs. In this work we provide a brief overview of time series analysis in general and review the contributions of the matrix profile in the field. We present applicable sequential optimizations to the SCRIMP kernel, on which we base our implementation. In particular we examine two parallelization approaches: the first one is based on suggestions from literature. In contrast to it,
the second parallelization approach proposed by us in this work provides scalability to longer time series, as it is not bound by single-node hardware. To assess the quality of the implementations we perform a series of scaling experiments. We investigate potential scaling bottlenecks of the implementations and fit a runtime model to predict tractability of time series analysis when applying the approach in practice.

Übersetzter Abstract:
Time series data is naturally formed by repeated measurements over time and their analysis in various science and engineering disciplines precedes the digital age. With progressing digitalization the amount and detail of such data is ever more growing and drives the demand for performant data mining tools to gain insights from the increasing amounts of data. The matrix profile proposed is a data structure which can serve in a variety of time series data mining tasks like motif search or clustering. Such analysis can be performed with little effort given the matrix profile, but obtaining the latter is a computationally intensive task. In this work we investigate migration of the matrix profile computation to a high performance computing cluster utilizing the MPI standard. Highly optimized parallel routines of MPI and the vast number of available computing elements on such clusters allow users to scale their algorithms in terms of runtime and level of detail according to their needs within the most distant limits. We present an approach, which enables the analysis of previously intractably large time series based on the matrix profile. It provides nearly unlimited scalability to solve large problems within reasonable runtimes according to the users needs. In this work we provide a brief overview of time series analysis in general and review the contributions of the matrix profile in the field. We present applicable sequential optimizations to the SCRIMP kernel, on which we base our implementation. In particular we examine two parallelization approaches: the first one is based on suggestions from literature. In contrast to it, the second parallelization approach proposed by us in this work provides scalability to longer time series, as it is not bound by single-node hardware. To assess the quality of the implementations we perform a series of scaling experiments. We investigate potential scaling bottlenecks of the implementations and fit a runtime model to predict tractability of time series analysis when applying the approach in practice.