



# Themis Palpanas

(Paris Descartes University)

## 「Scalable Machine Learning on Large Sequence Collections」

2020年1月9日(木) 10時30分～12時00分

東京大学本郷キャンパス理学部1号館9階 913号室

There is an increasingly pressing need, by several applications in diverse domains, for developing techniques able to analyze very large collections of sequences, or data series. Examples of such applications come from astrophysics and a multitude of other scientific and application domains that need to apply machine learning techniques for knowledge extraction. It is not unusual for these applications to involve numbers of data series in the order of hundreds of millions to billions, which are often times not analyzed in their full detail due to their sheer size. However, no existing data management solution (such as relational databases, column stores, array databases, and time series management systems) can offer native support for sequences and the corresponding operators necessary for complex analytics.

In this talk, we argue for the need to study the theory and foundations for sequence management of big data sequences, and to build corresponding systems that will enable scalable management and analytics of very large sequence collections. We describe recent efforts in designing techniques for indexing and analyzing truly massive collections of data series that will enable scientists to run complex analytics on their data. Finally, we present our vision for the future in big sequence management research, including the promising directions in terms of storage, distributed processing, and query benchmarks.