

# Machine Learning for Data Series Indexing

## Mots clés :

- **Directeur de thèse** : Themis PALPANAS
- **Co-encadrant(s)** :
  - Ioana Ileana
- **Unité de recherche** : Laboratoire d'Informatique PARIS DEscartes
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Sciences de l'information et de la communication

## Résumé du projet de recherche (Langue 1)

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). 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 index, ADS+: 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. 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 examine how machine learning techniques can be used in order to enhance the functionality of data series indexing, make them more efficient, and enable selectivity estimation and query answering cost estimation.