Title: Development of automated segmentation and clustering methods for spICP-ToF-MS time-series in Nanogeochemistry.
Supervisors: Mickaël Tharaud[1], Léonard Seydoux[1], and Themis Palpanas[2].
Level: Master 2 internship
Contact: Mickaël Tharaud (tharaud@ipgp.fr). Interested candidates should contact us by email with a resume and cover letter detailing their interest in the topic, relevant skills and academic background. Interviews will be arranged with shortlisted candidates to discuss their qualifications and fit with the project in more detail. Feel free to contact us with questions before submitting your application.

Working environment: This internship is funded by the data intelligence institute of Paris (diiP). The project will take place at the institut de physique du globe de Paris (IPGP), in collaboration with the laboratoire d'informatique of Paris Descartes (LIPADE). It will be supervised by Mickaël Tharaud, research engineer in analytical chemistry at IPGP, Léonard Seydoux, professor of seismology at IPGP, and Themis Palpanas, professor of computer science at LIPADE. The successful candidate will evolved in both environments, with a presence in both laboratories that will be defined according to the evolution of the research project.

Motivations and objectives: Nanoparticles (NPs) are pervasive in natural systems, playing a crucial role in nanogeochemistry. The emergence of single-particle time-of-flight inductively coupled plasma mass spectrometry (spICP-ToF-MS) has revolutionized NP characterization, presenting new challenges in data analysis. This research project seeks to bridge advanced nano-instrumentation with data-driven insights, focusing on the development of standardized methodologies for integrating spICP-ToF-MS with state-of-the-art machine learning algorithms. The IPGP hosts a world-leading geochemistry platform (PARI) equipped with an operational spICP-ToF-MS instrument and possesses an extensive dataset to (1) develop a novel methodology for the automated segmentation and clustering of NP time series generated by spICP-ToF-MS, (2) address challenges including instrumental noise, unknown NP compositions, and large data volumes requiring sophisticated statistical methods, and (3) explore interdisciplinary collaboration between geochemists, data scientists, and analytical chemists.

Proposed methods: Preliminary tests have shown encouraging results using a 4-step methodology described and illustrated below:

1. Detection: Establish a conservative threshold for detecting significant NP signals within time series data using intensity distribution across channels (b).
2. Clustering: Identify families of NP signals through unsupervised clustering algorithms, considering the unknown number of NP families in natural environments.
3. Classification: Train a classifier to differentiate various NPs within continuous time series, including an additional noise class, using realistic data.
4. Segmentation: Divide time series into segments based on the classifier, addressing the challenge of determining optimal segmentation window size (c).

Implementation Strategy: We will explore several solutions of each stage of the workflow onto several datasets including benchmarks, to allow us to select the optimal metaparameters. We will also systematically compare results with other methods mentioned in the state-of-the-art section and consider a deep learning approach using a denoising autoencoder if the proposed methodology proves unsatisfactory, contingent on the availability of sufficient synthetic data for training. If successful, the project aims to provide a ready-to-use API for automated segmentation and clustering of spICP-ToF-MS time series, with online accessibility and detailed documentation. This initiative will catalyze the development of new analytical methods for spICP-ToF-MS data, offering insights into the origin and fate of NPs in natural systems. The generic nature of the proposed method makes it applicable to various scientific domains dealing with time series data, potentially impacting fields such as seismology, chemistry, biology, and geodesy.

[1]Institut de physique du globe de Paris, Université Paris Cité, 1 Rue Jussieu, 75005 Paris
[2]Laboratoire d'Informatique Paris Descartes, Université Paris Cité, 45 Rue des Saints-Pères, 75006 Paris