Machine learning, a family of algorithmic methods that is at the core of artificial intelligence, allows a computer to learn a model that can be complex without explicit knowledge of its properties. While machine learning techniques have many potentialities, they also show many limitations. By means of two case studies in the field of planetary science, this postdoctoral project aims to investigate how machine learning algorithms can be appropriately constructed and used in the case when the training data is scarce or sparse (“data shortage”).

The first case study aims to advance the field of supervised machine learning by developing an architecture of transfer learning in the presence of sparse data. Supervised machine learning algorithms (e.g., neural networks) are typically trained “in isolation”, that is, each new instance of a classification or regression problem lacks the capacity to exploit information from previous experience. An example in planetary science is the recent development of neural networks for planet formation, which allows quickly and accurately predicting how planets form. However,
these networks lack the knowledge of how smaller bodies (e.g., asteroid) form. This limits their applicability. The candidate will be asked to experiment with the use of transfer learning to train new networks for asteroid formation, and to test if the network performances can be progressively improved by comparison with existing (but sparse) simulations.

The second case study aims to advance the field of unsupervised machine learning by developing \textit{new metrics for clustering analysis} of weakly correlated data. Unsupervised machine learning algorithms are not able to cluster data which are sparse and/or affected by statistical noise. An example in planetary science is the use of Hierarchical Clustering Methods (HCM) to search for asteroids originated in collisions, which show affine astronomical properties. However, the HCM is not able to detect clusters of asteroids if the collision happened too early during solar system history, because their data are weakly correlated. The candidate will be asked to develop new metrics which can be used to detect these weak correlations, and to progressively refine such metrics by using numerical and analytical simulations of asteroid orbital evolution.

The new techniques developed as part of this project have the potential to advance the science of machine learning by shedding lights on how to deal with (numerical and observational) data shortage. Importantly, this work also wants to highlight which are the limits of these algorithms. This will benefit not only the field of planetary science, but also other fields (such as the research in rare diseases or the collection of data for smart cities development) which are also characterized by lack of data.