On Capturing and Using Provenance in Machine Learning Pipelines

1. **Context**.

Machine learning pipelines are designed to generate predictive models given some raw data. Learned models are then utilized to make predictions given some (unseen) observations. The predictive power of the learned model depends largely on the data sets used for trained and how they have been preprocessed (engineered). ML-pipeline developers tend to rely mainly on their skills, past experience, and an iterative try-and-fail process to refine and improve ML.

2. **Objective**.

We seek to investigate how provenance information can be utilized to improve the process whereby ML-pipelines are designed and refined. In particular, the sub-tasks of the internships are as follows:

T1. A sweep of the state-of-the-art of provenance in data preprocessing and machine learning.

T2. Identifying techniques for the collection and utilization of provenance with the view to assist ML developers in the task of designing, improving, and debugging ML pipelines.

T3. The implementation of a prototype, and it is validation in the context of real-world ML pipeline.

3. **Work environment**.

The internship will last for 5 to 6 months, starting April the 1st 2021, and will be jointly supervised by Khalid Belhajjame (kbelhajj@gmail.com) and Daniela Grigori (daniela.grigori@lamsade.dauphine.fr) from the Paris Dauphine University. Technology-wise, we will be using mainly Python.

4. **Candidate**

The student must be a master student or an engineering student in his/her final year of study. To apply, send your CV, a letter of motivation and transcripts of the last three years to kbelhajj@gmail.com and daniela.grigori@lamsade.dauphine.fr.