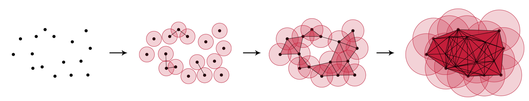
# A topological pipeline for machine learning methods

Francesco Conti - francesco.conti@phd.unipi.it

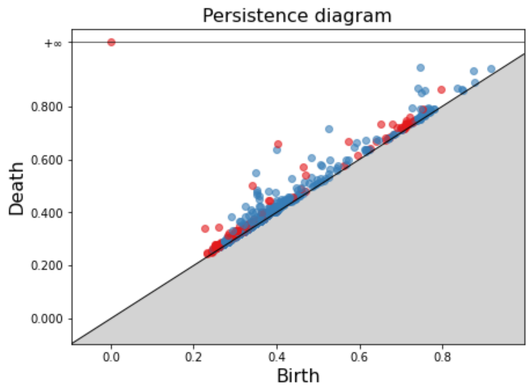
University of Pisa - Institute of Information Science and Technologies “A. Faedo”, National Research Council of Italy (CNR)

## Introduction

Topological Data Analysis (TDA) is a relatively new research field that aims to study digital data by means of topology. The main concept of TDA is Persistent Homology (PH), which studies a topological space at different scales and keeps track of birth and death of topological features.



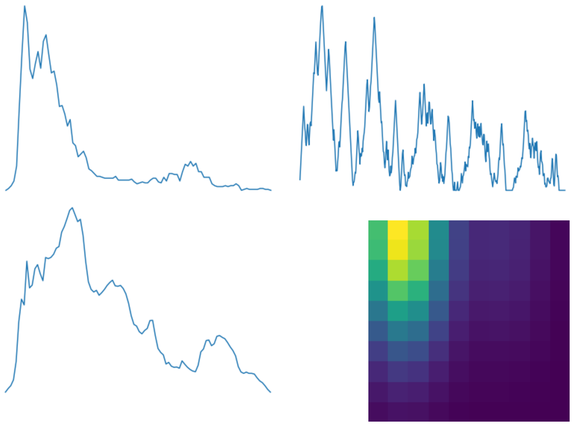
Data are assigned topological features by means of a filtration, which is a key aspect of the TDA paradigm. The collection of such features is called Persistence Diagram (PD).




Different filtrations yield different topological features, and thus different PDs.

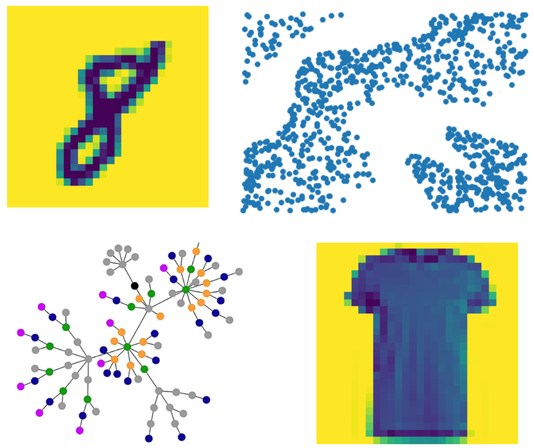
The choice of the right filtration is the first fundamental aspect of a TDA pipeline.

To introduce PDs in a machine learning algorithm they must be transformed as vectors. Multiple transformations are possible, and this choice is the second fundamental step in a TDA pipeline.



## Methodology

In this study we apply a topological pipeline to some datasets, in order to assess the quality of the various filtrations and vectorization methods with respect to the different types of data (i.e. images and dynamic trajectories).



## Preliminary results

The topological pipeline achieves excellent results in all the datasets. It consists in a novel way of study a classification problem for both benchmark and experimental datasets. In particular, the filtration strongly depends on the type of data and the dataset, while the vectorization method varies only between persistent images and persistent landscapes. The homology that achieves the best consistent results is the concatenation of both H0 and H1.

| Mean accuracy: | Dynamic | MNIST | FMINST |
|---|---|---|---|
| $H_0$ | $0.509 \pm 0.031$ | $0.915 \pm 0.006$ | $0.686 \pm 0.010$ |
| $H_1$ | $0.931 \pm 0.026$ | $0.617 \pm 0.007$ | $0.421 \pm 0.014$ |
| $H_0 + H_1$ fused | $0.565 \pm 0.048$ | $0.942 \pm 0.004$ | $0.716 \pm 0.013$ |
| $H_0 + H_1$ concat | $0.919 \pm 0.020$ | $0.940 \pm 0.006$ | $0.701 \pm 0.013$ |

| Best method: | Dynamic | MNIST | FMINST |
|---|---|---|---|
| $H_0$ | PI | PI | PL |
| $H_1$ | PL | PL | PL |
| $H_0 + H_1$ fused | PI | PI | PL |
| $H_0 + H_1$ concat | PL | PI | PL |

Also, a preliminary experiment was performed on a small experimental dataset, made of Raman spectra acquired on tissue samples for the diagnosis and grading of chondrogenic tumors. Best classification performances (binary classification) are quite promising: 98.4%, obtained on a 10-fold cross-validation, using the fused H0 and H1.

## Conclusion

This work is a first step towards both an easy, ready to use, pipeline for data classification using persistent homology and machine learning, and to understand the theoretical reasons why, given a dataset and a task to be performed, a pair (filtration, topological representation) is better than another. In the near future, we will apply this pipeline to other types of data, such as graphs, as well as experimental datasets coming from the ISTI laboratory.