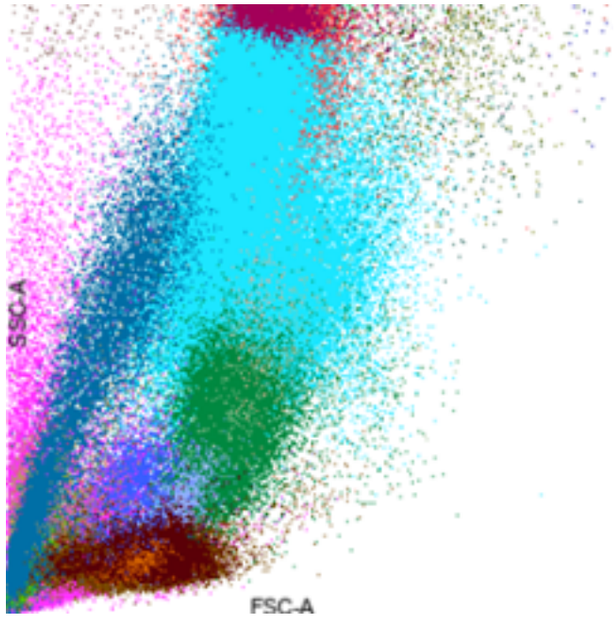
Kick-off chaires IA
Centrale Supelec - Sept 9, 2020

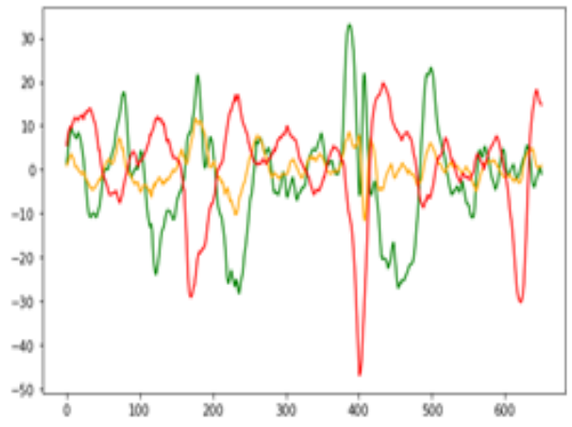
TopAI: Topological Data Analysis for Machine Learning and AI

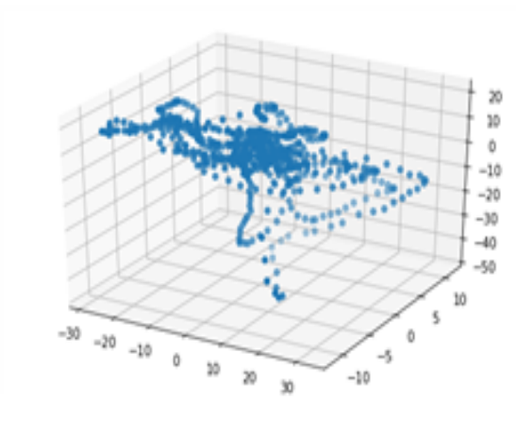
Frédéric Chazal
DataShape - INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr

What is Topological Data Analysis (TDA)?



[Cell population - cytometry - MetaFora courtesy]

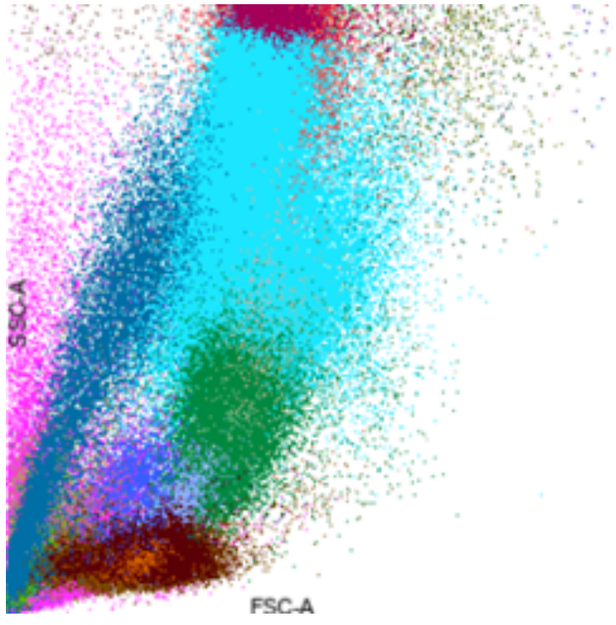




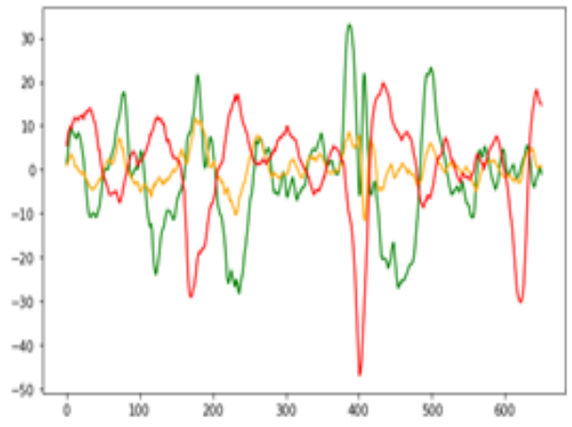
[Sensors (Sysnav courtesy)]

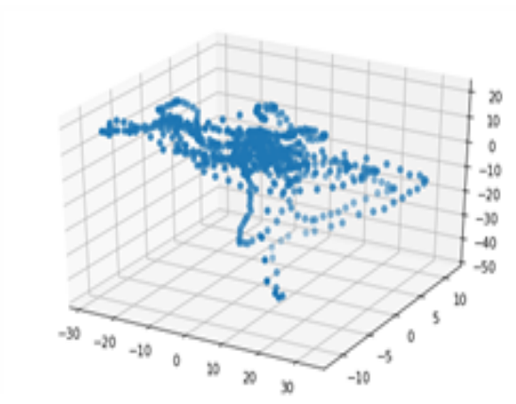
Modern data carry complex, but important, geometric/topological structure!

What is Topological Data Analysis (TDA)?



[Cell population - cytometry - MetaFora courtesy]



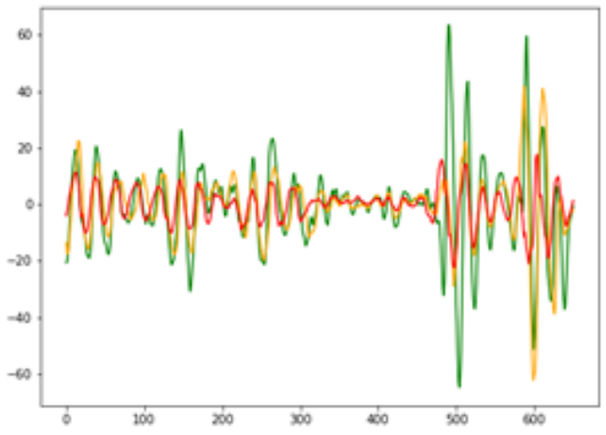


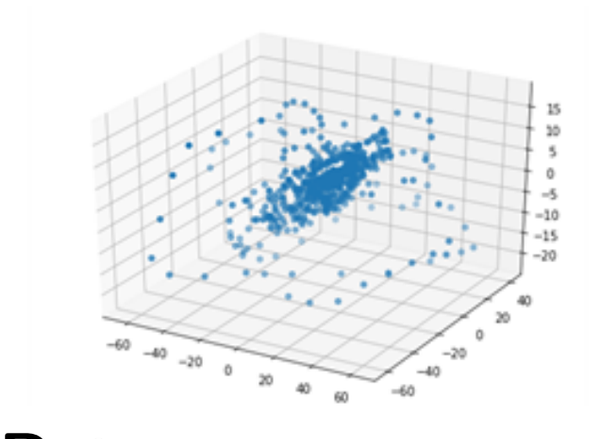
[Sensors (Sysnav courtesy)]

Topological Data Analysis (TDA) is a recent field whose aim is to:

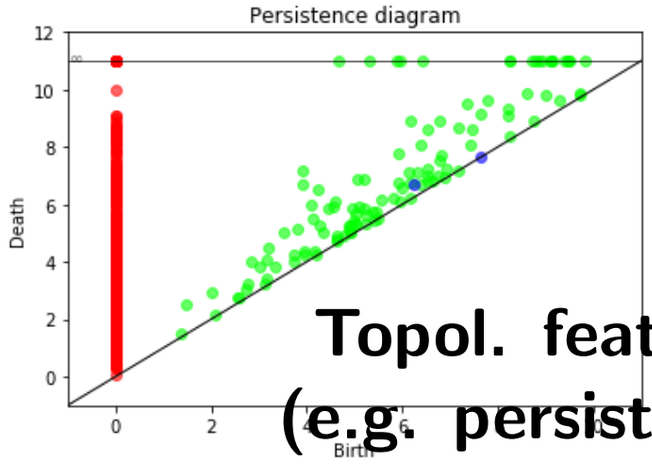
- infer relevant topological and geometric features from complex data,
- take advantage of topological/geometric information for further Data Analysis, Machine Learning and AI tasks:
  - using topological features in ML pipelines,
  - taking advantage of topological information to improve ML pipelines.

Topological information in ML pipelines





Data



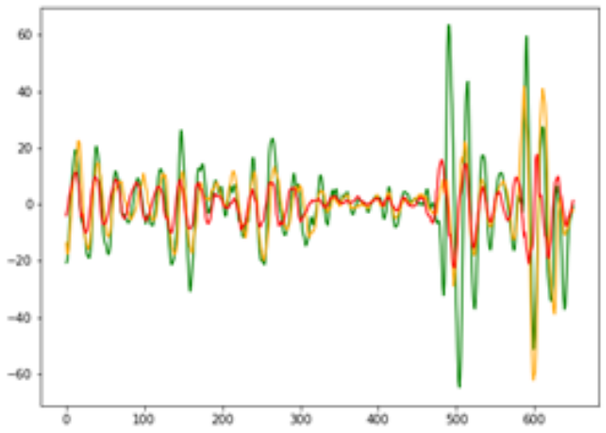
Topol. features (e.g. persistence)

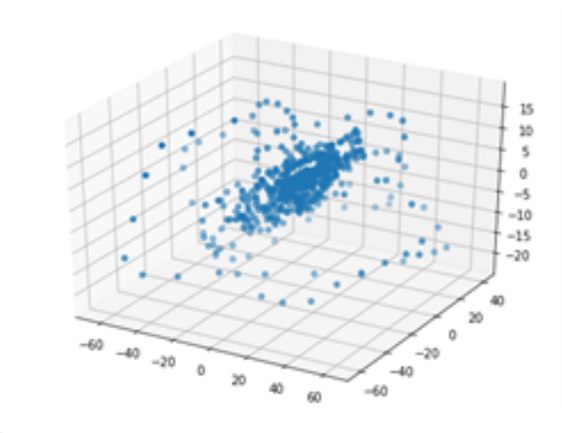
Strengths of topological information:

- robust, multiscale information,
- interpretable information (in some cases),
- benefit from a solid mathematical framework.

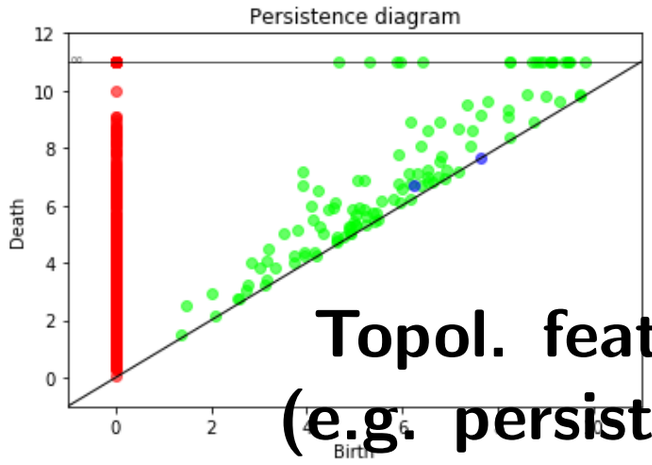
But...

Topological information in ML pipelines





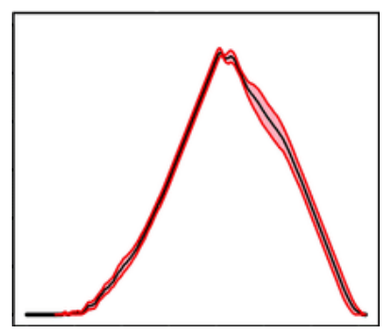
Data

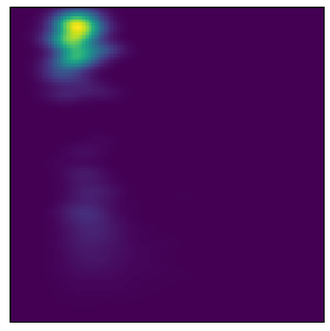


Topol. features (e.g. persistence)

- Topological information is usually not well-suited for classical ML algorithms (e.g. the space of PD is highly non linear)
- Not always clear which part of the topological features carries the relevant information.
- An active research area, but still at a very early stage (in particular regarding math. aspects).

Machine Learning / AI





Representations of persistence

The TopAI project

Overall objective: develop and promote the use of topological and geometric approaches in ML and AI, from mathematical foundations to industrial applications.

Two French industrial partners: sysnav NAVIGATION TECHNOLOGIES METAFORA biosystems

Four main axes:

- Mathematical foundations of TDA in ML.
- A generic software toolbox.
- Personalized medicine and diagnosis through clinical endpoints development.
- Topology-based unsupervised classification and anomaly detection on cytometry data for medical diagnosis.

informatics mathematics Inria

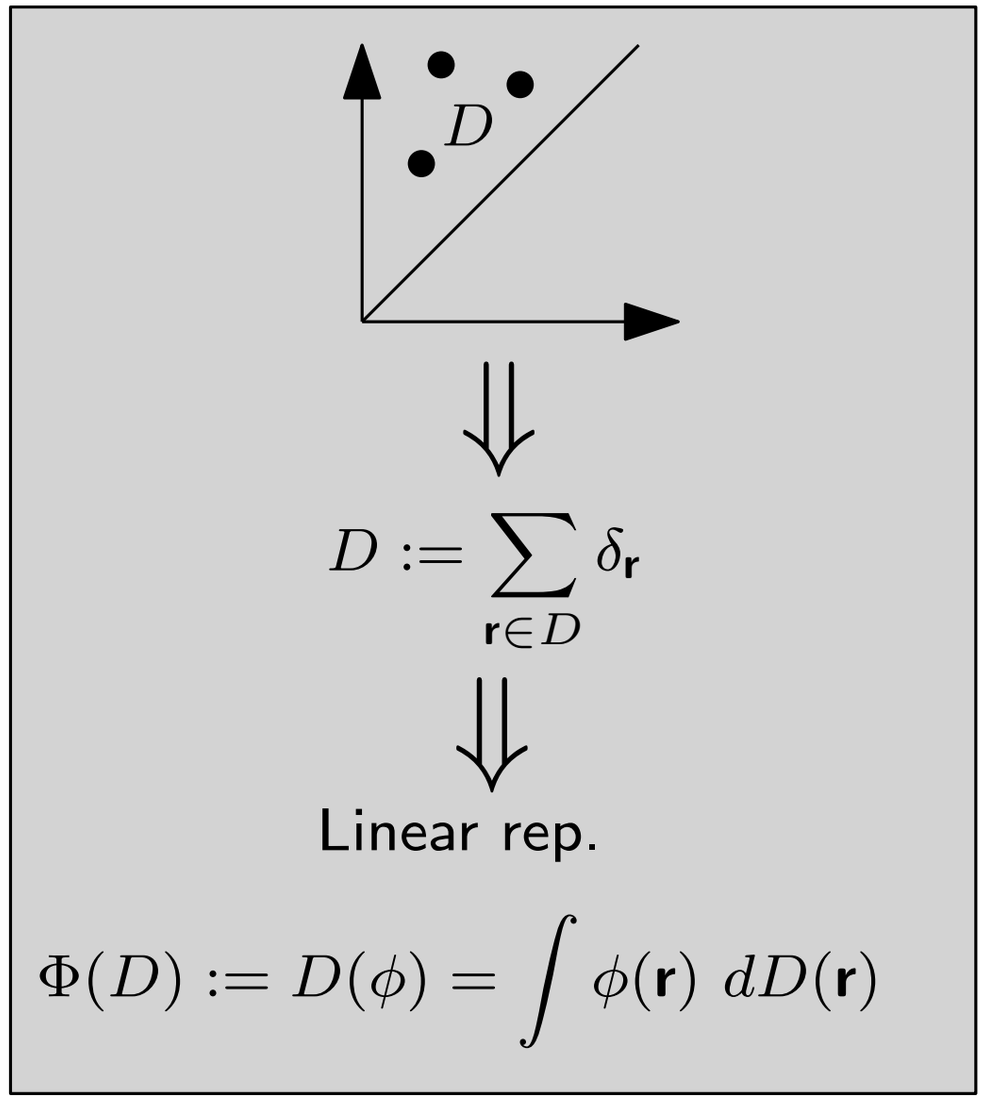
Mathematical foundations of TDA in ML

Main objective:

New TDA tools and TDA-based models coming with strong mathematical guarantees.

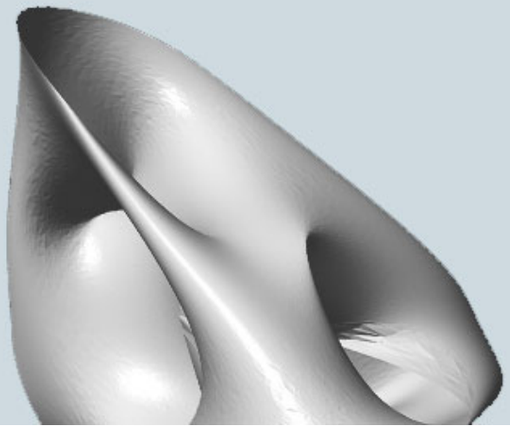
Methodology:

- TDA and (Deep) Learning:
  - learning topological features
  - exploiting topological structure of data distributions
- Statistical aspects of TDA:
  - parameter selection, topological losses and penalties
- Time-dependent and graph data
- Towards explainable AI

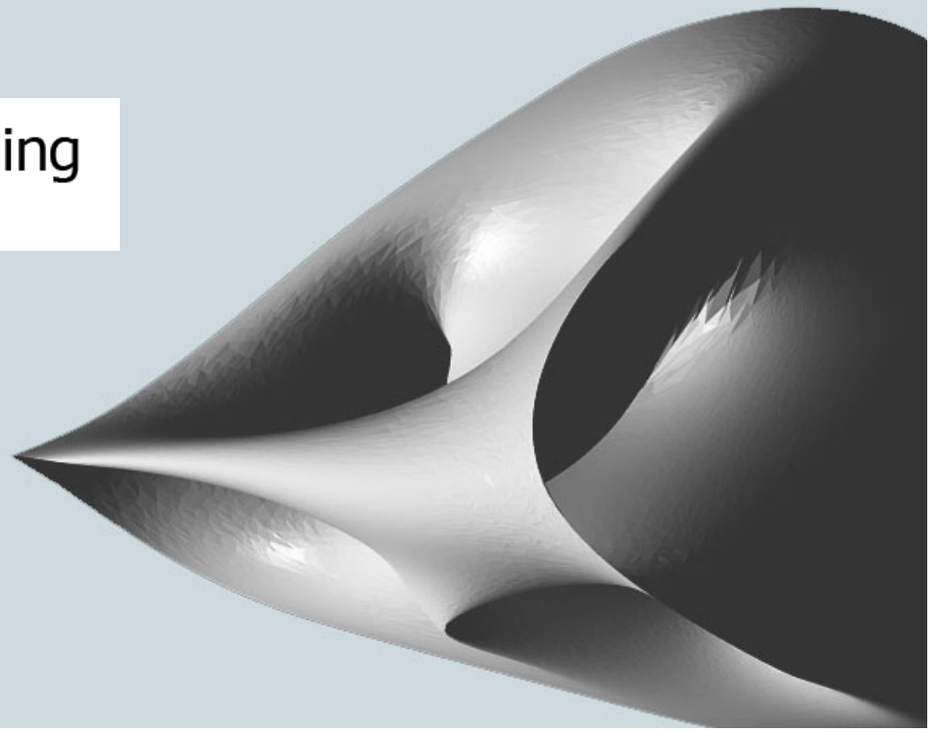


A generic software toolbox for TDA in ML and AI

गुढी GUDHI Geometry Understanding in Higher Dimensions



http://gudhi.gforge.inria.fr/



Main objective: Promote and transfer TDA towards non TDA experts data scientists and industry.

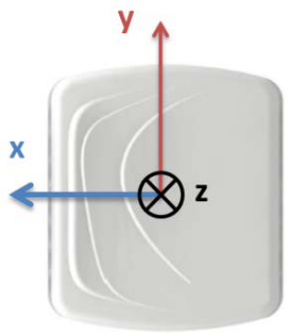
Methodology:

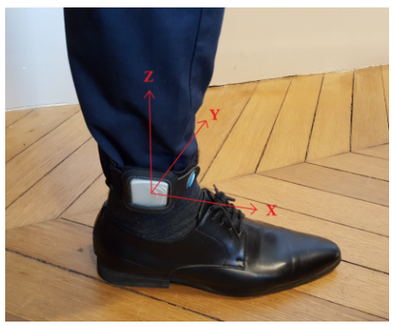
- Easy-to-use and efficient software implementing TDA methods for ML coming with relevant tutorials.
- Build upon the GUDHI libray that implements state-of-the-art data structures and algorithms for TDA.

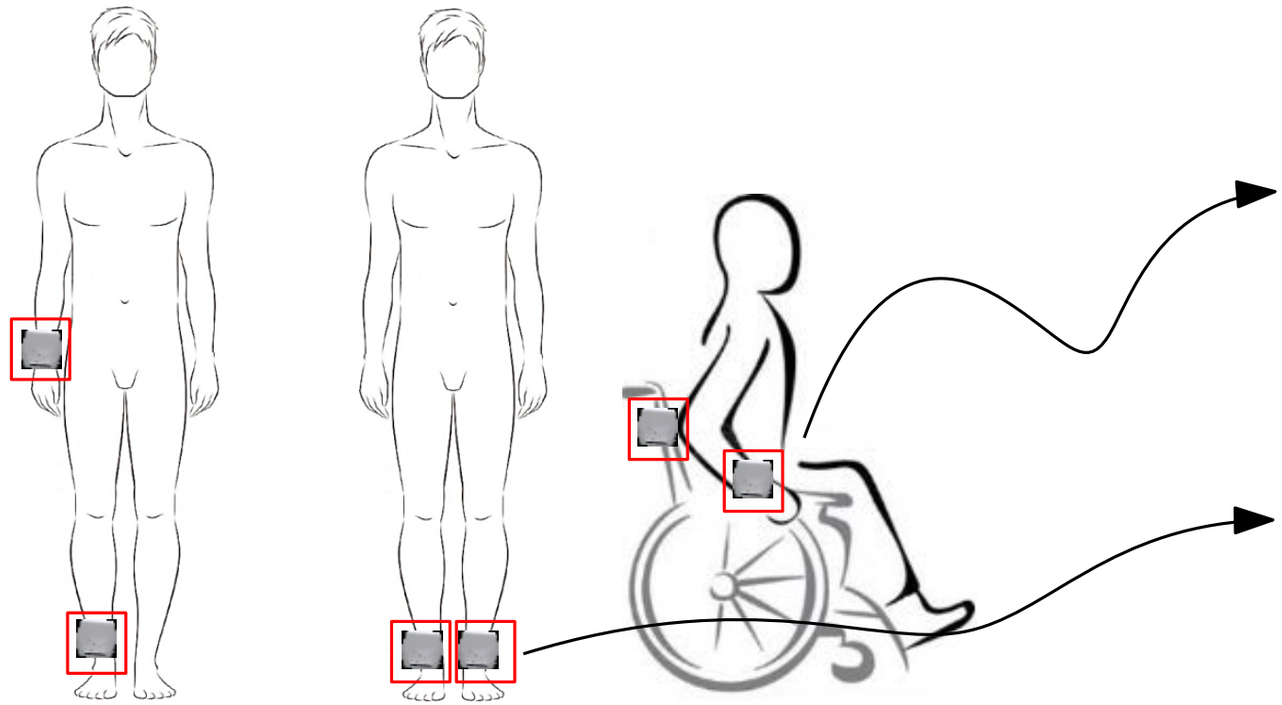
Personalized medicine and diagnosis through clinical endpoints development

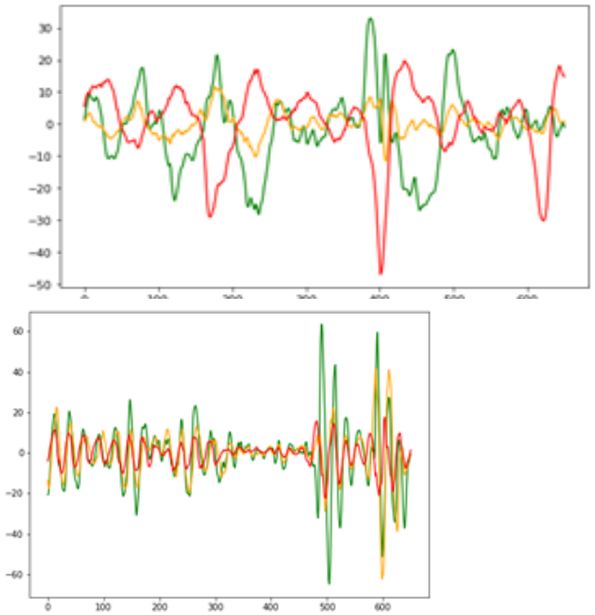
sysnav
NAVIGATION TECHNOLOGIES

A French SME with innovating technology to reconstruct pedestrian trajectories from inertial sensors (ActiMyo)









"Chaotic" time-dependent data

Objective: precise analysis of movements and activities of pedestrians.

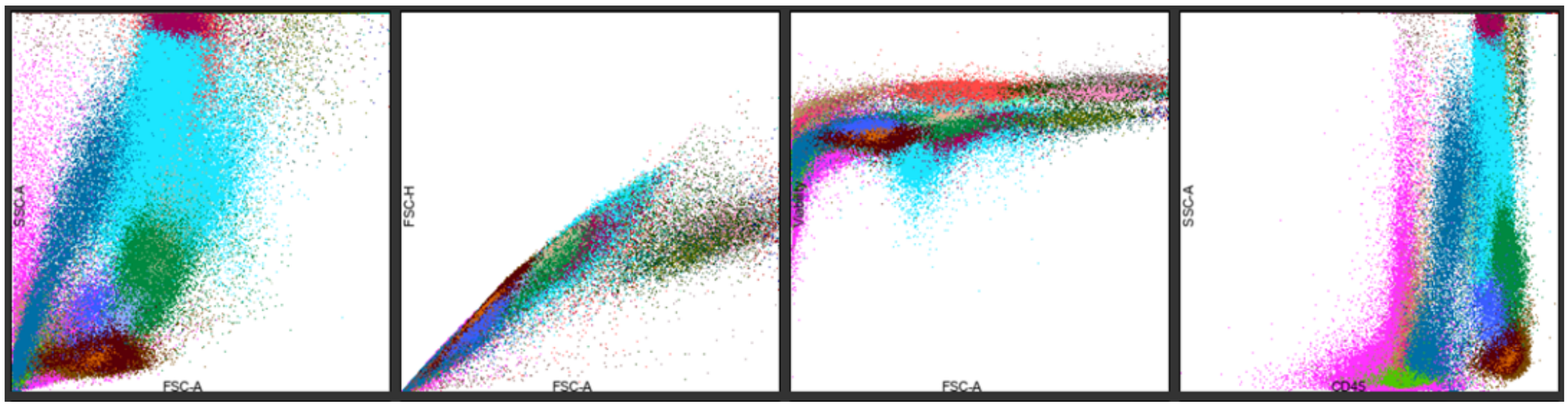
Applications: personal healthcare; medical studies; (defense).

Methodology: build on methods jointly developed by the partners and already successfully appled (EMA validation).

Topology-based unsupervised classification and anomaly detection on cytometry data for medical diagnosis

METAFORA biosystems

An innovative start-up specialized in biological diagnosis from cytometry data.



Objective: unsupervised learning in large point clouds (several millions) in medium/high dimensions ($\approx 4 \to 80$)

Applications: medical diagnosis from blood samples (1 point $=$ 1 blood cell)

Methodology: TDA based approaches, combined with dim. reduction methods to identify relevant patterns and subsamples.

Thank you for your attention!