Developing Topologically Informed Machine Learning Algorithms for Dynamic Data

Motivation: Inspired by my undergraduate research projects on machine learning (ML) and topological data analysis (TDA), I developed an interest in the intersection of the two fields, a rapidly growing area of research [5]. Persistent homology (PH), a powerful tool from TDA, captures the intrinsic structure of data. This information gives relevant topological input for ML models. While considerable research has been conducted at the intersection of the two fields in the case of static or time-embedded data, their intersection in the context of dynamic data remains relatively unexplored. Existing studies have used PH to extract topological features from time-varying data [6]. In addition, Kim and Mémoli have explored the theoretical aspects of dynamic PH [4]. Building on this foundation, I will focus on the computational aspects and **develop a multitude of innovative, optimized persistent homology and machine learning (PHML) pipelines in the context of dynamic data.** Such methods have the potential to facilitate predictions about the spatiotemporal evolution of data, an invaluable inquiry with numerous real-world applications.

Background: Consider data that evolves with time, such as a flock of birds, medical images as a tumor grows, or particles mixing in the surface layer of the ocean. At discrete time steps, PH can be applied to understand the "shape" of the data. In particular, PH captures how topological features– such as connected components, holes, and voids– change through a nested sequence of spaces, called a filtration [2]. Persistence diagrams, barcodes, and formigrams are used to summarize this information. When stacked across multiple time steps, these summary plots yield time-varying topological features, such as vineyards, persistence images, and CROCKER matrices. Persistence diagrams are not well suited to be inputted directly into ML models, so feature extraction methods, such as those in part D of Fig. 1, are a crucial step in the process. Notably, CROCKER matrices have already proven useful in some ML tasks [1, 6]. A vast array of unexplored dynamic PHML pipelines exist. **I seek to research these pipelines for their effectiveness in different tasks, considering computational costs and real-world applicability.**

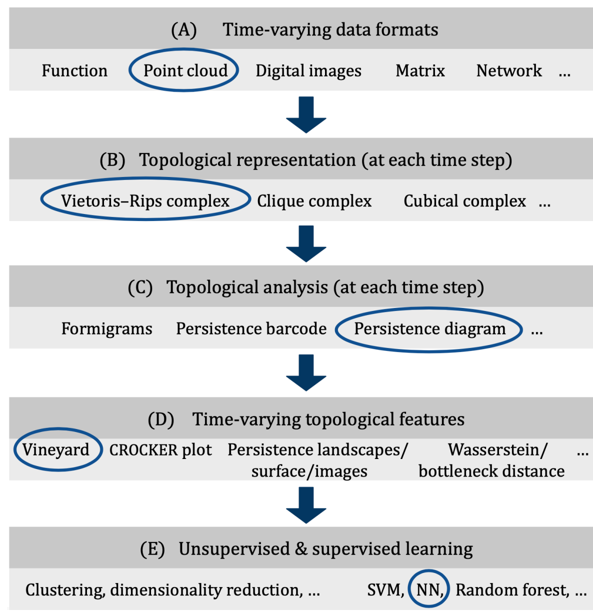

Figure 1: Flowchart for persistent homology and machine learning (PHML) algorithms in the context of dynamic data (adapted from [5]).

Goal 1: Assess the effectiveness of a given dynamic PHML pipeline to address specific tasks. Some dynamic PHML combinations may prove better at solving certain tasks (e.g. classification) than others. As a starting point, I will focus on the novel pipeline circled in Fig. 1. This pipeline involves the construction of persistence diagrams at each time step, stacked into a vineyard, vectorized, and subsequently fed into a neural network. The vectorization of a vineyard for the use in a neural network has yet to be done. It has been demonstrated that the weighted Vietoris–Rips simplicial complex effectively analyzes resource coverage [3]. With this new proposed methodology, one can address questions about the temporal evolution of coverage. While spatiotemporal graph-based neural networks have proven valuable in analyzing evolving patterns, the majority of studies rely on a classic representation of data [7]. My work can offer new ways to represent such data using topological structure. I seek to explore other combinations of tools from Fig. 1 that offer comparable potentials in the context of dynamic data.

Goal 2: Optimize the dynamic PHML pipelines. Optimizing the PH and ML steps to reduce computational costs is a vital aspect for the practical implementation of such approaches. Existing literature recognizes the high computational cost of generating some topological features [1]. I will experiment with methodologies, such as sampling and approximation strategies, to find innovative ways to save computation. Once the computational cost of the algorithm is reduced, I will implement it into a robust, publicly available package.

Goal 3: Apply and compare the performance of dynamic PHML with existing tools. Using a robust package developed in Goal 2, I will demonstrate its use in a wide range of applications. For example, the highlighted approach in Fig. 1 can be applied to study the spatiotemporal demand of energy. Important questions such as when and where energy should be distributed to meet demand can be addressed. I plan to compare the performance of this work to existing tools that forecast energy demand, potentially sparking collaboration with city-based energy suppliers or the U.S. Energy Information Administration. Given its versatility, dynamic PHML has the potential to address novel interdisciplinary questions about spatiotemporal data.

Intellectual Merit: This work presents an opportunity to provide researchers across diverse fields with innovative, efficient, and robust pipelines for analyzing dynamic phenomena. My proposed research provides numerous avenues to advance both PH and ML. With **strong research and academic foundations in these underlying computational fields,** I am well equipped to innovate PHML pipelines in the context of dynamic data. To further prepare myself for this line of research, I am currently investing my time in a graduate-level class on TDA to learn about cutting-edge tools in the field.

Broader Impacts: Given the quantity of unexplored dynamic PHML pipelines, there is an abundance of research directions, many of which are approachable to undergraduate researchers. Therefore, I will mentor undergraduates at my graduate university to explore the robustness of pipelines in real-world, time-varying applications. A project of this type will provide young researchers with exposure to many computational methodologies and challenge them to think in interdisciplinary ways. In summary, the **multitude of dynamic PHML packages** I develop through this work will **make the data-driven research on time-varying phenomena more accessible, inspire novel investigations for both entry-level and experienced researchers, and foster interdisciplinary collaboration.**

References:
[1] Bhaskar, D. et al. **(2019)**. *Chaos an Interdisciplinary Journal of Nonlinear Science*, *29*(12).
[2] Dey, T. & Wang, Y. **(2021)**. *Computational Topology for Data Analysis*. Cambridge University Press.
[3] Hickok, A. et al. **(2024)**. *SIAM Review, 66,* 481-500.
[4] Kim, W. & Mémoli, F. **(2020)**. *SpringerLink*, *66*, 831–875.
[5] Pun, C. S. et al. **(2022)**. *SpringerLink*, *55*, 5169–5213.
[6] Xian, L. et al. **(2021)**. *Foundations of Data Science*, *4*(1), 1.
[7] Zeghina, A. et al. **(2024)**. *Neurocomputing*, *594*, 127861.