Data is ubiquitous in today's society. To interpret and to understand the data requires both technical theory to identify important features of the data and application expertise to interpret those features. My research is a collaborative and interdisciplinary approach using topological data analysis (TDA).

We can use TDA to summarize data in various ways. One way is to use persistent homology, which is (informally) a way to describe the components, tunnels, and voids of the underlying set from which the data was sampled. However, using persistent homology to directly compare data sets or to recognize meaningful features can be cumbersome: a problem for which I use statistics to overcome. I also leverage my theoretical results and apply persistent homology theory to road network analysis and to the grading of prostate cancer. I have made novel contributions to my research community and have published in premiere peer-reviewed venues. As I progress in my career, I plan to continue both developing the theoretical foundations and employing the theory in practical applications.

Dissertation Work: As a graduate student, I proved an inequality that bounds the difference between lengths of curves by a function of the Fréchet distance between the curves and the total curvatures of the curves. This result culminated in a single-author paper [11]. In another project, I studied Gaussian mixtures, a widely used – but poorly understood – data model. The existence of so-called *ghost modes* was first shown in [5]; however, in [10], we fully analyzed one case in which exactly one more mode than kernel appears. Over the past four years, my research has evolved to be more data-driven, but the desire to understand fundamental properties remains at the heart of most of my research.

The Intersection of Statistics and Computational Topology: One of the current challenges in persistent homology is to identify the pertinent topological descriptors of a data set. I investigate how statistics can enhance data analysis in TDA. Along with statisticians at CMU, I have defined and developed algorithms to compute confidence sets for persistence diagrams using techniques such as the bootstrap [6, 7, 12] , reaching a broad audience, both in the statistics and the theoretical computer science communities. We have also developed methods for subsampling data to compute stable topological descriptors of large data sets [8]. Along with Fabrizio Lecci and Jisu Kim, I created an R package that implements our statistical methods, making these tools accessible to other researchers.[1] Currently, we are investigating how to compute the power of various hypothesis tests in TDA. This work will be driven by applications in astronomy, comparing simulations versus observations of the distribution of matter throughout the universe [9].

Topological Analysis of Networks: Digital road maps/networks are an invaluable resource today, and much effort goes into keeping these maps current. A desirable distance measure between road networks is needed. Prior to my recent publications [1, 2, 3], the popular distance measures were mostly heuristic in nature and lack theoretical guarantees. I provide theoretical guarantees by explicitly using the embeddings of the maps.

The research community has been receptive to these ideas, as evidenced by invitations to give talks at workshops and universities, and a recently funded NSF grant focusing on analyzing data on road networks. In this project, I am currently working with an undergraduate, Maia Grudzien, in order to compare car accident data across cities.

Data Descriptors in Pathology: The widespread availability of digital pathology images opens up new possibilities to use computational approaches to leverage the information inherent within them for diagnosis, prognosis, and precision medicine. A recent collaboration with researchers at Tulane University aims to discover new quantitative image-based prognostic biomarkers (data descriptors) for prostate cancer, focusing on an investigation of novel concepts from TDA applied to prostate cancer glandular architecture. Funded by an NSF-NIH grant, we have subitted a proof-of-concept paper of our approaches [4], developed a collaborative annotation tool for working with a pathologist, and will be submittin two follow-up grants (one to NSF and one to NIH).

[1] `http://cran.r-project.org/web/packages/TDA/index.html`

References

[1] Ahmed, M., Fasy, B. T., Hickmann, K., and Wenk, C. A path-based distance for street map comparison, 2015. To appear, Trans. Spatial Alg. Sys. (TSAS) 2015. Preprint available at arXiv:1309.6131.

[2] Ahmed, M., Fasy, B. T., and Wenk, C. Local persistent homology based distance between maps. In *SIGSPATIAL* (Nov. 2014), ACM.

[3] Ahmed, M., Fasy, B. T., and Wenk, C. New techniques in road network comparison. In *Grace Hopper Celebr. Women Comput.* (Oct. 2014). Online proceedings.

[4] Berry, E., Brown, J. Q., Fasy, B. T., Lawson, P., and Wenk, C. Topological descriptors for quantitative prostate cancer morphology analysis. In submission.

[5] Carreira-Perpinán, M., and Williams, C. An isotropic Gaussian mixture can have more modes than components. Informatics Research Report EDI-INF-RR-0185, Institute for Adaptive and Neural Computation, University of Edinbugh, Dec. 2003.

[6] Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems 20*, 6 (2013), 96–105. Also available at arXiv:1311.0376.

[7] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Robust topological inference: Distance-to-a-measure and kernel distance, 2014. In submission. Preprint available at arXiv:1412.7197.

[8] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Subsampling methods for persistent homology, 2014. ICML. Preprint available at arXiv:1406.1901.

[9] Cisewski, J., Fasy, B. T., Hellwing, W., Lovell, M., Rinaldo, A., Wasserman, L., and Wu, M. Topological hypothesis tests for the large-scale structure of the universe. Work in progress.

[10] Edelsbrunner, H., Fasy, B. T., and Rote, G. Add isotropic Gaussian mixtures at own risk: More and more resilient modes in higher dimensions. *Discrete Comput. Geom.* (Jun. 2013), 797–822.

[11] Fasy, B. T. The difference of length in curves in $\mathbb{R}^n$. *Acta Sci. Math. (Szeged) 77* (2011), 359–367.

[12] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *Annals of Statistics 42*, 6 (2014), 2301–39. Preprint available at ArXiv:1303.7117.