MATH – BIOINF – STATS 547: Mathematics of Data

Winter 2025 (3 credit hrs.): **Open to upper-level undergraduates and graduate students**
Instructor: Prof. INDIKA Rajapakse (E-mail: indikar@umich.edu, Phone: 734-615-3134)
Teaching Assistant: MARC Andrew Choi (E-mail: machoi@umich.edu)
Class Time: Tuesday and Thursday, 2:30 - 4:00 PM,
Office Hours: Wednesday and Friday (INDIKA R), 3:00 PM - 5:00 PM: https://meet.google.com/dnm-zipr-tsb or **in person after class**, Wednesday 2:00 PM - 3:00 PM (Zoom) and Thursday 1:00 PM - 2:00 PM in person at East Hall Atrium (MARC C)

Prerequisites: This course is open to graduate students and upper-level undergraduates in applied mathematics, bioinformatics, statistics, and engineering, who are interested in learning from data. Students with other backgrounds such as life sciences are also welcome, provided they have maturity in mathematics. The mathematical content in this course will be linear algebra, multilinear algebra, dynamical systems, and information theory. This content is required to understand some common algorithms in data science.

Description: I will start with a very basic introduction to data representation as vectors, matrices, and tensors [1,2]. Then I will teach geometric methods for dimension reduction, also known as manifold learning (e.g. diffusion maps, t-distributed stochastic neighbor embedding (t-SNE), UMAP, etc.), and topological data reduction (introduction to computational homology groups [3], etc.). I will bring an application-based approach to spectral graph theory, addressing the combinatorial meaning of eigenvalues and eigenvectors of their associated graph matrices and extensions to hypergraphs via tensors. I will also provide an introduction to the application of dynamical systems theory to data including Dynamic Mode Decomposition. Finally, I will present an introduction to Transformers [4]. Real data examples will be given where possible and I will work with you to write code implementing these algorithms to solve these problems. The methods discussed in this class are shown primarily for biological data, but are useful in handling data across many fields. This course features several guest lectures from industry and government.

Workload: We will have five homework assignments and a final project in lieu of a final exam.

Resources: A comprehensive website containing all reading materials and class notes will be maintained throughout the term.

Premise: This class is motivated by my own experience with data and mathematics, and also inspired by Dr. Steve Smale's work [4]. Over the past few years I have worked with Dr. Smale and with my funding institute DARPA on many problems that have required knowledge at the interface of mathematics and data. The DARPA initiatives on Artificial Intelligence, the Lifelong Learning Machines (L2M) and Guaranteeing AI Robustness Against Deception (GARD) programs, exemplify the types of work at this interface. My own lab has a dual approach of generating time series data in-house and using mathematics to identify patterns in data and determine major unknowns. The methods discussed in this class reflect our approaches and those that are useful in handling data across many fields.

References:

1. Strang G. (2019). Linear Algebra and Learning from Data (2019). Wellesley-Cambridge Press
2. Eldén L. (2015) Matrix methods in data mining and pattern recognition. Society for Industrial and Applied Mathematics.
3. Edelsbrunner H., Harer J. (2010). Computational topology: an introduction. American Mathematical Soc.
4. Bishop C. M., Bishop H. (2023). Deep learning: Foundations and concepts. Springer Nature.

These books are not required reading, but you may find them useful if you want a second opinion or more detail on a topic