SYLLABUS: GEOMETRIC DATA ANALYSIS

1. Overview

The goal of this class is to introduce approaches to analyzing data presented as finite metric spaces using ideas from algebraic topology and differential geometry. Prerequisites are a grounding in basic probability, statistics, and linear algebra. The class will focus on rigorous mathematical foundations and applications drawn from computational genomics.

Evaluation.

- Students will be assigned weekly problem sets; each problem set will have both theoretical and programming components. The programming assignments will be based on running examples from genomic data sets.
- There will be a take-home midterm, again incorporating both theoretical and implementation components.
- There will be a final project; the students will be encouraged to work in teams for the final project. Satisfactory final projects might include:
  - Novel data analysis using existing methods.
  - A theoretical study of a mild refinement of an existing algorithm.
  - A careful review of a relevant paper not covered in class.
  - Systemic analysis of the performance of a method covered in class on data that violates its hypotheses.

2. Topics

(1) **Clustering. (1 week)**
- A rapid review of basic clustering algorithms (e.g., hierarchical clustering, k-means, relaxation approaches to k-means, spectral clustering).
- A discussion of axiomatic approaches to reasoning about the possible performance of clustering algorithms (notably Kleinberg and Carlsson-Memoli); a brief introduction to category theory and functoriality.

(2) **Manifold learning and related topics. (3 weeks)**
- The basics of differential topology (i.e., what is a manifold, what is a Riemannian manifold),
- An overview of techniques in manifold learning (LLE, Laplacian and Hessian eigenmaps, manifold charting, dimensionality estimation)
- Non-manifold approaches to obtaining coordinates on nonlinear objects (t-SNE, UMAP) using local geometry
- Heat kernel/diffusion approaches for multiscale analysis (Coifman, Maggioni), including a brief introduction to spectral graph theory.

(3) **Metric geometry. (2 weeks)**
- Basics of metric geometry, curvature and CAT(k) spaces, cubical complexes (Burago, Villani).

- Gromov-Hausdorff distance and Gromov-Hausdorff convergence, with discussion of the space of phylogenetic trees (Billera-Holmes-Vogtmann) as a motivating example.

(4) **Metric measure spaces and optimal transport. (2 weeks)**
- An introduction to metric measure spaces, weak convergence of probability distributions, (Gromov)-Prohorov and (Gromov)-Wasserstein distances.
- Statistics on non-positively curved metric measure spaces (Sturm).
- Matching metric measure spaces.

(5) **Topological data analysis. (4 weeks)**
Topics include:
- An introduction to simplicial complexes, homology and homotopy.
- Persistent homology and stability theorem for persistent homology.
- Hardness results for topological inference (Weinberger).
- Topological machine learning (e.g., embeddings of topological features in Banach spaces).