PROJECT IDEAS FOR TOPOLOGICAL DATA ANALYSIS AND MACHINE LEARNING

1. A list of implemented ideas

This section contains hands-on project ideas, mostly from the academic domain. All these are implemented projects, so they will be useful to get a feel of an end-to-end pipeline.

1.1. **Applications to problems in pure mathematics.** There are works in the literature where the data/objects come from pure mathematical structures, and topological data analysis techniques (especially persistent homology and its generalizations) are used to study them. Below are some references where TDA is applied to data derived from mathematical objects.

- Persistent Homology of Configuration Spaces of Trees
- Persistent Path Homology of Directed Networks
- Data-Driven Perspectives on Knot Invariants
- Structure of the chromatic polynomial
- Big data approach to Kazhdan–Lusztig polynomials

1.2. **Protein structure classification.** *Classify protein structures (e.g., alpha vs. beta types) using persistent homology features derived from 3D point clouds representing atomic coordinates* Use publicly available protein datasets. Project includes extracting persistent barcodes and vectorizing them for use in classifiers.

- Persistent homology reveals strong phylogenetic signal in 3D protein structure
- Mathematical Insights into Protein Architecture: Persistent Homology and Machine Learning Applied to the Flagellar Motor.

1.3. **Social network community detection.** *Can persistent homology help classify different types of 'communities' or structures (e.g., tightly-knit vs. dispersed groups) in social networks?"* Use public data to predict or classify according to sociological labels.

- Community Detection via Persistent Homology Surfaces.
- Community detection, pattern recognition, and hypergraph-based learning: approaches using metric geometry and persistent homology

1.4. **Time Series Anomaly Detection in Sensors.** *Apply persistent homology to rolling windows of time series from IoT sensors to detect anomalies (e.g., machine failures or ECG abnormalities).* Generate sliding window embeddings, compute persistent diagrams from delay embeddings, and use these for anomaly prediction. Detect periodic vs. non-periodic patterns in physiological time series (e.g., ECG, respiration) using sliding window embedding + persistent homology.

- Passive encrypted IoT device fingerprinting with persistent homology.
- Anomaly Detection Using Persistent Homology.
- Time Series Classification via Topological Data Analysis.
- Machine learning of time series data using persistent homology
- Exact multi-parameter persistent homology of time-series data.

1.5. **Shape Classification in Images (Medical or Industrial).** *Classify images (e.g., cancerous vs. normal cells, manufactured vs. defective parts) using topological features extracted by persistent homology.* Employ pre-processing (contour extraction or binarization), then compute topological signatures and apply vectorization (persistence images/landscapes).

- Persistent Homology Guided Medical Image Classification
- Texture image classification based on persistent homology.

1.6. **Crystal Structure Identification.** *Use persistent homology to distinguish between different crystalline structures in simulated materials or microscopy images.* Extract point clouds, analyze the topology, vectorize the persistence diagrams, and classify the material type.

- Topological representations of crystalline compounds for the machine-learning prediction of materials properties.

1.7. **Authorship Attribution.** *Can persistent homology on word co-occurrence graphs (or syntactic dependency graphs) help attribute authorship in literary texts?* Build graphs for documents, compute persistent homology, and use vectorized summaries for classification.

- Topology-aware Authorship Attribution of Deepfake Texts.

1.8. **Market Regime Prediction.** *Does persistent homology on point cloud embeddings of financial time series help in predicting market regime shifts (bull vs bear)?* Generate time-delay embeddings, analyze with persistent homology, and vectorize results for time series classification.

- Application of Persistent Homology in Forecasting Realized Volatility
- A persistent-homology-based turbulence index & applications in financial markets.
- A persistent-homology-based turbulence index & some applications of TDA on financial markets
- Topological tail dependence: Evidence from forecasting realized volatility Author links open overlay panel

1.9. **Applications in text analytics.** *Can topology offer new insight in text embeddings?* Search for topological features in text - an entire document or word embeddings. Look for novel ideas that would study text from shape view-point.

- A Novel Method of Extracting Topological Features from Word Embeddings.
- Persistence Homology of TEDtalk: Do Sentence Embeddings Have a Topological Shape?
- A survey of TDA applications in NLP.

2. A list of potential research projects in pure mathematics and industry

All these are high-level ideas with potential, but possibly not so well-defined. You may have to figure out how exactly to make TDA work in the given context. Feel free to make appropriate changes to the problem statement or how should TDA apply.

2.1. **Knots and Links.** The objects are knots and links, represented via large tabulated datasets. Each knot comes equipped with numerous computable invariants:

- crossing number,
- Alexander, Jones, and HOMFLY polynomials,
- signature,
- genus,
- hyperbolic volume (when applicable),

- fiberedness, alternatingness, etc.

These invariants embed knots into a high-dimensional numerical space. Following are the sources to collect the data.

- **KnotInfo**: `https://knotinfo.org/homelinks/about.html`
- **Knot Atlas**: `http://katlas.org/wiki/Main_Page`
- SageMath Knots Documentation.

TDA methodology: We treat each knot as a point in $\mathbb{R}^d$, where coordinates are normalized invariants. Two complementary approaches are natural:

(1) **Mapper**: reveals families of knots and transitions between classes (e.g. alternating $\leftrightarrow$ non-alternating).
(2) **Persistent Homology**: detects clustering and higher-order features in the space of invariants.

Typical lens functions include:

- crossing number,
- hyperbolic volume,
- first principal component of invariant vectors.

Some of the research questions one could explore.

- Do fibered knots form distinct topological clusters?
- Are there persistent cycles corresponding to mutation classes?
- Can Mapper graphs predict knot properties not used as features?

2.2. **Low-dimensional Topology.** The study of low-dimensional topology has produced extensive censuses of triangulated and hyperbolic 3–manifolds. Each manifold admits computable invariants such as:

- hyperbolic volume,
- number of cusps,
- first homology group,
- length spectrum approximations,
- triangulation complexity.

The existing databases are:

- **SnapPy manifold census**: `https://github.com/3-manifolds/SnapPy`
- **SnapPy documentation**: `https://snappy.math.uic.edu`
- **Regina triangulation software**: `https://regina-normal.github.io`

Some questions for exploration could be:

- Are there topological "holes" in the census suggesting missing manifolds?
- How would you interpret persistent 1-cycles?
- Can clusters predict geometric properties not used as coordinates?

2.3. **Elliptic Curves and Arithmetic Geometry.** Elliptic curves over $\mathbb{Q}$ form one of the richest arithmetic datasets available. Each curve is associated with invariants such as:

- conductor,
- Mordell–Weil rank,
- torsion subgroup,
- Tamagawa numbers,
- analytic invariants.

Some of the existing databases are:

- **LMFDB**: `https://www.lmfdb.org/EllipticCurve/Q`
- **Cremona tables (via Sage)**: built-in database
- LMFDB project overview: `https://www.lmfdb.org/about`

Each elliptic curve is a point in a high-dimensional arithmetic feature space. TDA can be used to:

- detect clusters corresponding to isogeny classes,
- identify loops caused by congruence conditions,
- study "geography" of ranks and torsion.

Some more research questions for exploration are:

- Do rank distributions exhibit topological stratification?
- Are torsion structures separable via Mapper?
- Can PH detect any "anomalies"? How would one interpret them?

2.4. **Customer Analytics & Marketing.**

- **Problem**: Segment customers more robustly than k-means or PCA clustering.
- **TDA idea** Apply Mapper or PH on embeddings of purchase histories / browsing patterns to detect "hidden" subgroups of customers (e.g., seasonal vs habitual buyers).
- **ML Task**: Churn prediction, recommender system improvement.
- **Novelty**: Topology captures global relationships between customer types that clustering misses.

2.5. **Financial Market Anomaly Detection.**

- **Problem**: Anticipate market crashes or regime shifts.
- **TDA Idea**: Use sliding-window persistent homology on financial time series (stocks, forex) to detect "shape changes" in correlation networks before events.
- **ML Task**: Early warning classification (stable vs unstable regime).
- **Novelty**: TDA detects structural market shifts not visible in volatility measures alone.

2.6. **Supply Chain & Logistics Bottleneck Detection.**

- **Problem**: Identify choke points in logistics networks.
- **TDA Idea**: Model supply chain as a weighted graph, apply PH to cycle structure (loops = redundancies, bottlenecks = missing cycles).
- **ML Task**: Predict delivery delays, optimize routing.
- **Novelty**: Topology provides structural resilience measures beyond graph centrality.

2.7. **Predictive Maintenance.**

- **Problem**: Detect machine failures before downtime.
- **TDA Idea**: Apply PH to sensor time series (vibration, temperature) → persistent cycles as fingerprints of normal vs faulty operation.
- **ML Task**: Fault classification, RUL (Remaining Useful Life) prediction.
- **Novelty**: PH features are noise-robust, outperforming raw signal features.

2.8. **Cybersecurity Intrusion Detection.**

- **Problem**: Detect cyberattacks in real-time from network traffic logs.
- **TDA Idea**: Apply PH to graphs of connections / flows; unusual loops or holes may indicate malicious activity.
- **ML Task**: Binary/multi-class intrusion classification.
- **Novelty**: TDA provides topological fingerprints of anomalous traffic.

2.9. **Smart Energy Grid Resilience.**

- **Problem**: Detect anomalies in energy demand or grid failures.
- **TDA Idea**: Apply PH to time series of energy usage across smart meters; persistent features reveal unusual spikes or breakdowns.
- **ML Task**: Anomaly detection, demand forecasting.
- **Novelty**: TDA robustly detects nonlinear anomalies (e.g., load shedding, rolling blackouts).

2.10. **Text Analytics for Customer Feedback.**

- **Problem**: Extract themes from customer reviews, support tickets.
- **TDA Idea**: Apply PH on word embedding spaces (Word2Vec, BERT) to cluster recurring complaint types.
- **ML Task**: Topic detection, sentiment classification.
- **Novelty**: TDA captures semantic geometry of complaints better than LDA clustering.