This talk will be about data-driven questions and how topology can help with some of them.

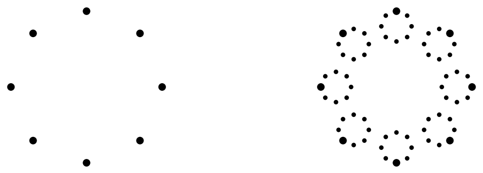
Goal of TDA: Given a point sample, can we:
- predict where the next sample will be?
- explain any "patterns" in the given sample?

Informally, it is a "what will hapen" vs a "why did something happen" approach.

Remark: We know everything about "Euclidean space," or $\mathbf{R}^N$, but very little about other spaces in general. Most methods try to interpret the difficult spaces as $\mathbf{R}^N$, or parts of $\mathbf{R}^N$.

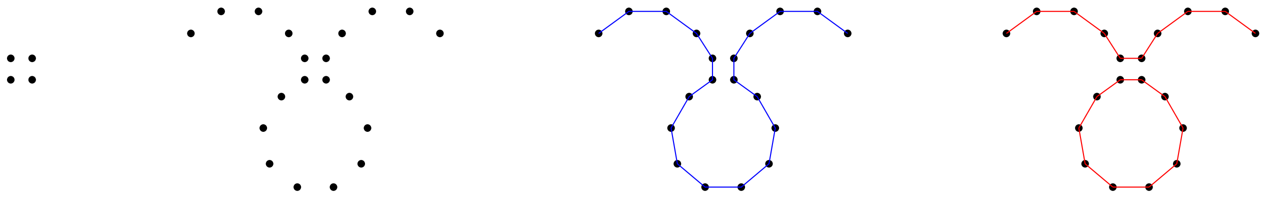
0.1 Scaling problems

Suppose we are given sample on left. Where will next sample be? Now suppose given next sample on right. Where will third sample be? Which was the "correct" prediction?



Remark: "Correctness" depends on the scale chosen. The problem becomes: which is the correct scale?

Sometimes even the problem is at the same scale. Suppose we begin with small sample on left, then expand to sample on right. How many pieces does original shape have? One or two?

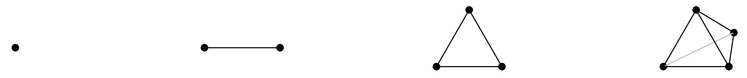


This particular problem can be made easier by:
- Sampling more points (may not always be possible)
- Knowing more about the shape (like how curved it is)

example: facebook (maybe not)

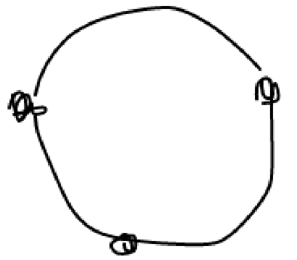
0.2 Persistent homology

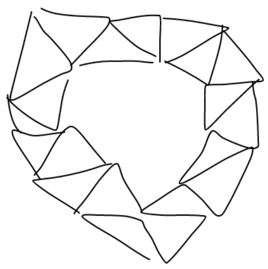
Goal: Create a topological space from a point sample (so usual tools can be applied). All (nice) topological spaces can be decomposed / approximated by putting together *simplices*:



0-dimensional 1-dimensional 2-dimensional 3-dimensional

They are the bulding blocks in topology. For example, a circle and annulus can be made from them, as below. Note that we always match up the points, edges, and faces when putting together the pieces. The collection of simplices is called a *simplicial complex.*





0.2.1 The Čech complex

Let $P = \{p_1, \dots, p_n\}$ be a point sample of elements $p_i \in \mathbf{R}^N$, and $t \in \mathbf{R}_{>0}$. This is also a (very uninteresting) simplicial complex made of 0-dimensional simplices. We will make it more interesting in the following way.

1. Put balls of radius t around every point $p_i \in P$.
2. If $B(p_i, t) \cap B(p_j, t) \neq \emptyset$, add a 1-simplex that attaches to p_i and p_j.
3. If $B(p_i, t) \cap B(p_j, t) \cap B(p_k, t) \neq \emptyset$, add a 2-simplex that attaches to p_i, p_j, p_k.
4. Keep going like this up to n-simplices.

Result: Analyze space with usual topological tools. Things to look for / what they imply:

How many separate pieces are there?	How many distinct groups were sampled from?
What is the (maximum) dimension of the space?	How simple / redundant is the data?
How many different ways to get from p_i to p_j?	How many unique relationships between p_i and p_j?

Example: Points in plane.

Extension: The n-sphere S^n is viewed by topologists as the the most fundamental shape. Most questions can be boiled down to:

How similar to a sphere is the shape / a piece of the shape?

More precisely, topologists study all the functions $f : S^n \to X$, compare them, classify shapes by them. This is called the *homology* of the shape. Functions into simplicial complexes are very easy to understand, so "toplogical classification" is very easy!

Problem: What radius t do we choose? What is the right scale?
Solution: Choose all radii. More specifically:

1. Choose a very small lower bound a, a very large upper bound b, and split up $[a, b]$ into s pieces.
2. Calculate $\check{C}(P, a + k\frac{b-a}{s})$ for $k = 0, \dots, s$.
3. Analyze particular feature of space at every radius t.
4. The features that *persist* for many t can be considered to be properties of original space.

Example: Points in plane. When radius too small, nothing new. When radius too big, all info lost.

Real world example: Ezra Millers wings:
- Different patterns of veins in fruit flies' wings, some closed loops, some not
- Reflect "evolutionary distance" between species of fruit flies

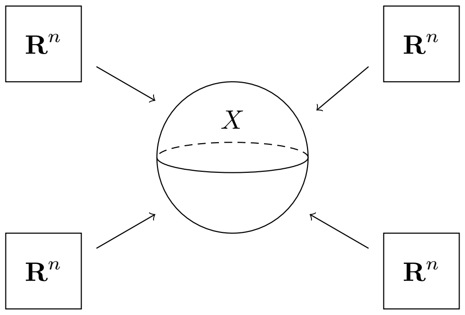
Not quite "sample → complex → interpretation", rather "observation → future data interpretation".

0.3 Sampling / statistical techniques

Reverse question: Given a very complicated shape (or topological space) called X, can we use a less complicated shape that retains all the properties of the original?

Applications: This is useful to train computers to understand known shapes:
- Testing for accuracy
- Machine learning

Definition: A topological space X is an n-dimensional *manifold* if there exists a collection of (smooth) functions $\{f_i : \mathbf{R}^n \to X\}$ whose images cover X.



Very complete description of X, but too many points. Instead we want a *uniform sample.*

Facts: We know how to do the following things very well:
- Sample uniformly in $\mathbf{R}^n$
  - Every subset of $\mathbf{R}^n$ that has the same size is equally likely to have a point sampled from it
- Calculate minimum number of uniformly sampled points whose Čech complex is $\cong X$
- Calculate derivatives and integrals

Problem: Sampling uniformly on $\mathbf{R}^n$ may not be uniform on X when taken through f_i.
Solution: Calculate integral / derivative on $\mathbf{R}^n$, adjust sampling function
- Number is integral of Jacobian (determinant of matrix of second derivatives) on paramaterizing domain

---

0.4 Dimensionality

Current problem: How can higher-dimensional topological information be interpreted? Most care about "connectedness"

Rare case: The evolution of viruses was thought of as a tree. It has been shown that recombination of viruses happens, creating closed loops.
- TDA tools can analyze huge amounts of data to find "loops"
- Gives a new language to talk about evolution

Open question: Where can 2-dimensional data be used? What is a "function of S^2 into a space" mean?

0.5 Further reading

Computational Topology: An Introduction by Herbert Edelsbrunner and John Harer (book)

Topology and Data by Gunnar Carlsson (article)

Finding the Homology of Submanifolds with High Confidence from Random Samples by Partha Niyogi, Stephen Smale, and Shmuel Weinberger (article)

Data Structures for Real Multiparameter Persistence Modules by Ezra Millert (arXiv preprint)