BMC - Advanced: *Applied* Algebraic Topopology

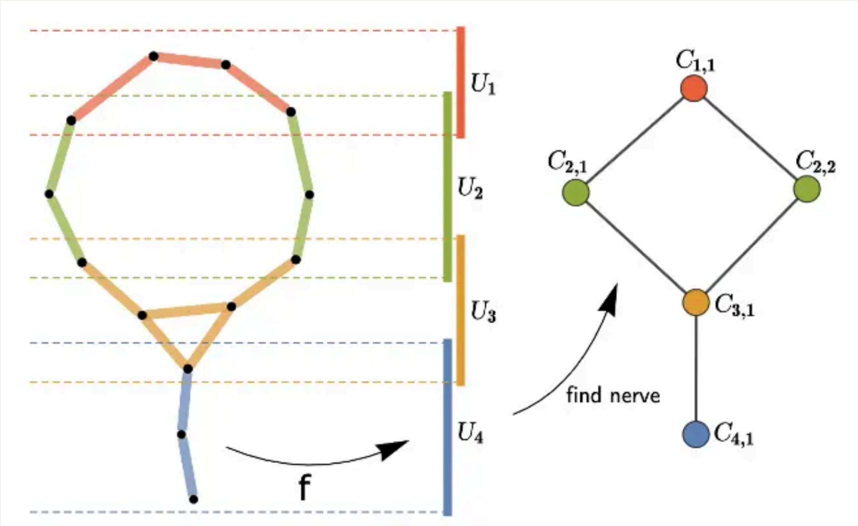
220309 - Chris Overton (handout based on revision after class)

The second talk on 220316 will recap and expand on these notes.

Caution: these notes are not written to be fully understandable to you on their own, although they will be filled in a bit from class presentation. We try to cover details in class, and you are encouraged to look up terms!

When you see a paragraph marked "question" (or Q), try to answer the question before looking past <---SPOILER ALERT---> at the answer.

The "mapper" algorithm (preview):

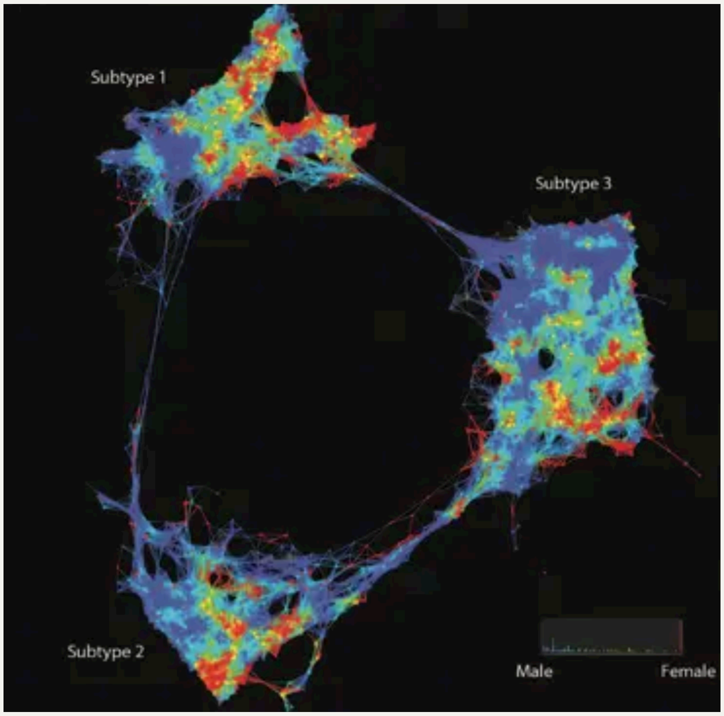


***Applied* Algebraic topology not only uses Algebraic Topology ("AlgTop"), but it uses it in "real world" applications, like data science.**
One very active area of research is **Topological Data Analysis ("TDA")**

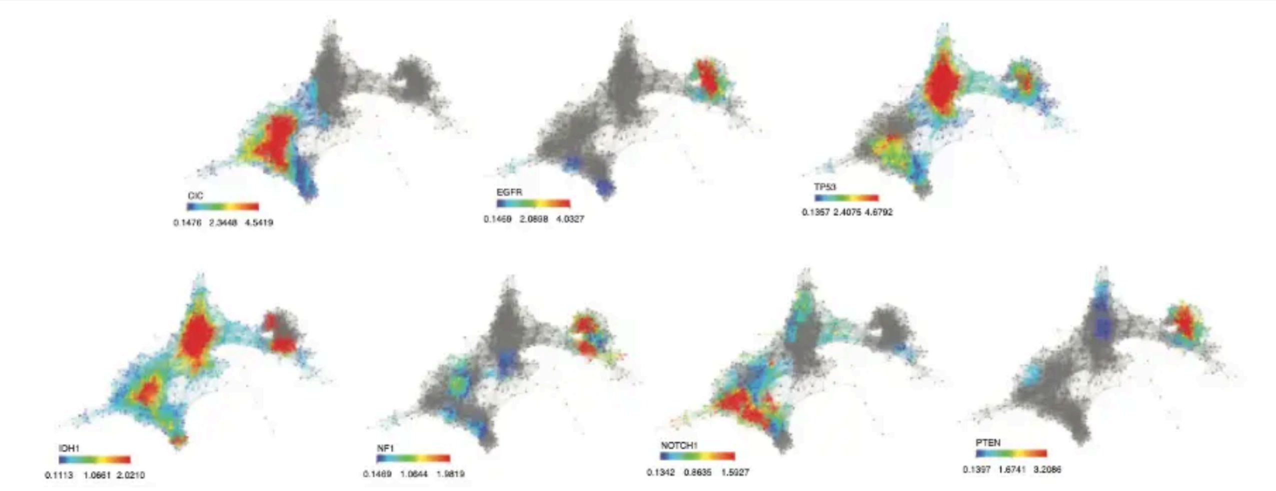
Alg Top is one of the major areas of "pure" math. You use it to map from topological spaces to objects you can study using modern algebra.

Understanding TDA requires not only AlgTop (itself tricky!), but then you have to connect this say to data science, which has its own background.

Examples of how you can understand a data set using the mapper algo:
1: Diabetes: discovering innate structure of a complex data set



2: Cancer: discovered geometric structure may link with causal characteristics - here mutated genes



Our plan for day 1:

- Introduction to the mapper algo, and how it linked seemingly pure AlgTop with not just real world applications, but interesting Silicon Valley startup history
- Gaining some respect & intuition for high-dimensional spaces - how can you even build them?
- Stepping back a bit to introduce necessary math ideas:

  Topology and AlgTop (we won't have much time for Alg):
    3 views of continuity
    Homotopy
    First view of AlgTop: homology
    Manifolds, a bit of geometry
- How TDA brings topology/geometry together with data science:
  Clustering, "dimension reduction"
  How can you do topology on a set of points?
    Nerves

Plan for week 2:

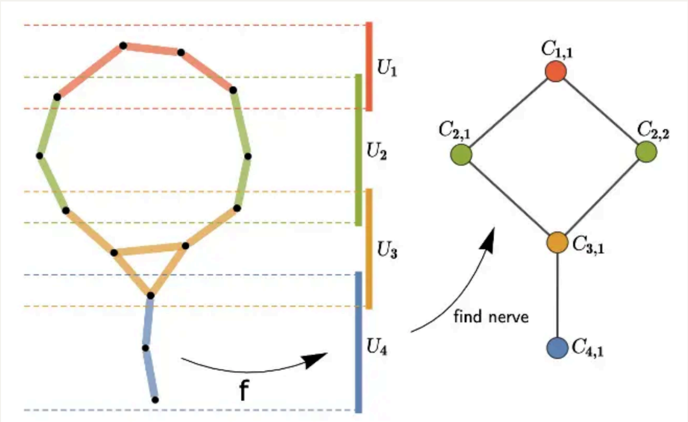
- A more careful review of homology
- More TDA examples, tools, and techniques:
  e.g. persistent homology
- More credible "computations"

The mapper algorithm

Goal: you have a cloud of points (maybe in high dimensional space) that you'd like to understand. Try to capture its structure in a graph (or simplicial complex)

Steps:

1. One or more "lenses": projection into low-dimensional space (e.g. a line)
2. "Cover" the image with overlapping intervals (or higher-dimensional sets)
3. Take preimages of sets from the cover; find "connected components" of these (really: cluster.) These make up vertices of your new graph
4. You have the underlying original points that end up in the clusters. If clusters in adjacent preimages share points, connect them (creating edges, maybe higher-dimensional objects)
5. If you have made good choices for lenses, cover, and clustering, the resulting graph will capture important structure!



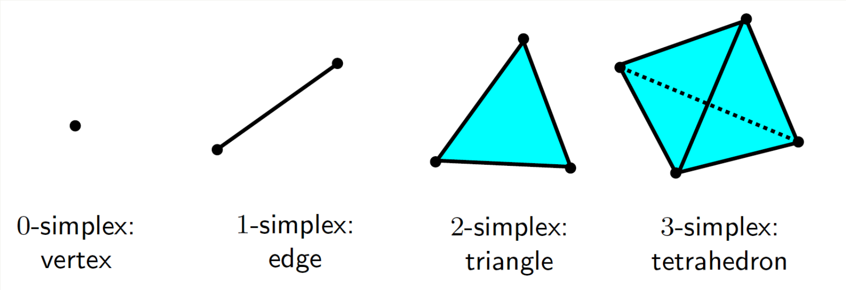
Historical context:

- '07 paper by Singh, Mémoli & Carlsson
- Topological/geometric analogue of "map reduce" - a very important algorithm in CS & technology
- Went from thesis project to company with >$100M funding (Ayasdi)
- That was later bought out privately and is now more quiescent
- Meanwhile, TDA has caught on as an appealing research area in math

Gaining some respect & intuition for high-dimensional spaces - how can you even build them?

Some possible building blocks:

- Hypercubes I_n
    - How many corners does an n-cube have?
- n-dimensional balls B_n with boundary S_{n-1} (a sphere)
- Simplexes (or simplices) Δ_n - "hyperpyramids":
    - How many k-faces does an n-simplex have?



Questions to consider:

Q1) Prove that each of $I_n(r), \Delta_n(r), B_n(r)$ has a volume with formula $c_n * r^n$, where r is a "radius" and c is different for each of these three kinds of shapes (and for each dimensionality n.)

Note: if r is not given, it is assumed to be 1, indicating shapes with unit edge length in each direction (or radius 1 for balls.)
(Sometimes the n are written as superscripts instead of subscripts.)

Q2) Which of the following statements are true? Do they contradict each other?

- Given a choice of $\epsilon > 0$, as n becomes large, "almost all" of the volume of B_n is in the slice of width ϵ closest to the equator.
- Given a choice of $\epsilon > 0$, as n becomes large, "almost all" of the volume of B_n is in the outer "peel" of thickness ϵ near the boundary.

<---- SPOILER ALERT---->

Answer 1) This is actually quite easy: calculate a given answer for r=1, then scale by a factor of r in each of n dimensions (geometric similarity.)

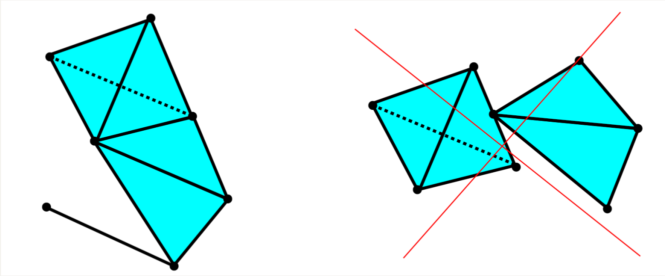
Answer 2) Surprisingly, both seemingly contradictory statements are true.

To prove the first, use Pythagoras' theorem to note that each slice i of say M total slices parallel to the "equatorial hyperplane" other than the central one has maximum radius $r_i \lneq r$. Using the similarity result from Q1, this shrinks volume by at least the power $(r_i/r)^{n-2}$ (for all the n-2 other dimensions.) When n is sufficiently large, this gets smaller than any specified ϵ/M, so you can make the total volume of the M-1 non-central slices as relatively small as you want.

The second statement follows more directly from similarity (Q1.)

When you combine Q1 and Q2, it follows that as n gets sufficiently large, "almost all" the volume is concentrated on the "peel" of the central slice!

You saw that cubes have more complicated boundaries. So simplexes are easier to build spaces out of:



We just insist that simplices of the same dimension that overlap have to overlap in a common face-simplex. This is violated above on the right, where the vertex in the middle of the edge of the 3-simplex is not a vertex of the the 3-simplex.

Topology: some fundamental concepts

Continuous maps f: three views, the third leading to a "topology":

- Can you draw a graph of f without lifting up your pen?
  (Applies only to maps from 1-d to 1-d)
- $\epsilon - \delta$ definition: by holding your "gun" within δ precision, you can hit a "target" of radius ϵ
  (Applies only to metric spaces: spaces with distances)
- Topological: inverse $f^{-1}(U)$ of an open set U is open

A topology of a space X is nothing more than a specified set of "open" subsets $\{U_i \subset X\}$:

- Including $\emptyset$, X
- Closed under **arbitrary** unions
- Closed under **finite** intersections

Examples of topologies on the real unit interval I_1:

- $I_{1[t]}$: The "trivial topology": only $\emptyset$ and I_1 are open
- I_1: The "usual" topology - how could you define this?
- $I_{1[d]}$: The "discrete" topology: every subset is open

Question 3: given identity maps, which of the following are continuous?

- $I_{1[t]} \to I_{1[d]}$
- $I_{1[d]} \to I_{1[t]}$
- $I_1 \to I_1$

<---- SPOILER ALERT---->

Answer 3) The 2nd and 3rd are continuous, because any open set in the image has a pre-image (in this case the same set of points) that is open.

Remember: maps tend to be continuous if they do not "break apart" open sets.
The important thing to understand is that topological continuity depends on pre-images of the map: going backwards from the direction of the map!

What **maps** are allowed?
Acceptable maps must respect the structure of the topological spaces considered. In most cases, one would want them to be continuous. But if one is working with a differentiable **category*, the maps would also have to be differentiable.

*Note: we won't have much time to talk about categories, but they are very generally useful things that are basically like sets (**objects**) between any of two which there exist a set of **morphisms** (like maps) between them. Plus there are unique identity maps from an object to itself, and morphisms are associative.

You can also map from one category to another in a way that commutes with the objects and morphisms. This is called a **functor**. AlgTop studies functors from categories with topological objects to categories with algebraic objects.

Some important definitions in topology for a topological space X:

- An **open cover** of X is a set of open sets U_i whose union contains X
- X is **compact** if every open cover has a finite sub-cover
  A deep and important concept (another math circle topic I did!)

(Topological) Spaces considered equivalent:

- A consequence of the usual topology on manifolds: you can stretch spaces, but you can't puncture them or cut them.
- When are two spaces considered equivalent? Two possibilities:
  - Only if homeomorphic, i.e. if there is a bijection (*1-1* and *onto*, so an invertable map) between the spaces that is continuous in each direction
  - If homotopic, namely if you have maps going both ways whose compositions are homotopic to identities

Q4 Example: P and O (ignoring serifs!) are homotopic, but not homeomorphic. Why?

<---- SPOILER ALERT---->

One proof is that by removing one point of P, you can turn it into two (topologically separate) pieces. Any homeomorphism between P and Q would force such a separation of "P minus a point" to map onto a disconnected "Q minus a point", which is not possible.

Manifolds

Definition: an n-dimensional manifold is a topological space in which each point x has an open "neighborhood" N_i homeomorphic to the interior of B_n via a map $f_i : N_i \to X$. The f_i are called charts.
Also: if the images of two such charts f_i and f_j overlap, the sequence $f_j^{-1} \circ f_i : N_i \to N_j$ has to be **smooth** (i.e. it has to have a specified number of continuous derivatives.)

This lets you think of a complicated space just in terms of its charts, which are defined on subset of Euclidean space.

Similarly, the maps you allow between manifolds are only those that result in smooth maps between neighborhoods in chart.

Intro to AlgTop

There are surprisingly few Alg Top texts aimed at undergrads - most are for graduate students, maybe allowing in some "advanced undergraduates."

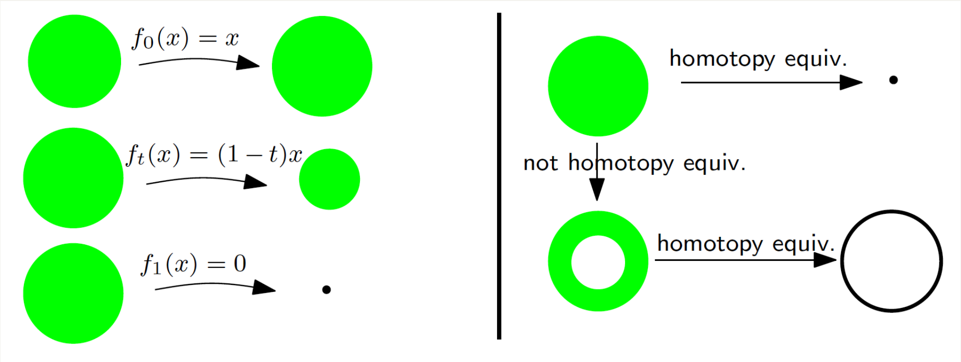
Why? Because you first need to know some things about 1) Algebra, 2) Topology, and then you combine these in new ways.

We have skimped on both (particularly on Algebra), *and* will also introduce AlgTop only very briefly.

What do you wish you knew?

One example might be: what are all the different possible connected spaces, and all the allowed maps between them (continuous or smooth, depending on which is your chosen category.)

There are way too many such spaces and maps, so instead, you consider smaller sets, such as homotopy equivalence classes of spaces



But even just homotopy types of spaces (and maps) are way too hard to compute. For example, we only know a very small portion of all the sets of homotopy classes of maps between spheres $S_m \to S_n$.

So a much more workable thing is **homology**:

Homology: (graded) groups formed e.g. by "chains" of simplexes of which X is composed

There are a couple ways to set up homology, ranging from the most concrete (build X out of simplexes) to definitions that are more abstract but more convenient.

Suppose X is itself a 'simplicial complex', namely composed of points, line segments, triangles, tetrahedra, hyper-tetrahedra, ... (respectively in dimensions 0,1, 2, 3, 4). An n-simplix is given by its n+1 vertices as $c_n = (v_0, v_1, \ldots, v_n)$. You can take 'chains' of these, which are sums of multiples of n-simplexes, forming the **module** C_n.

Then there is a boundary operator $\delta_n : C_n \to C_{n-1}$, defined on a simplex as:
$\delta_n(v_0, v_1, \ldots, v_n) = (v_1, \ldots, v_n) - (v_0, v_2, \ldots, v_n) + (v_0, v_1, v_3 \ldots, v_n) + \ldots$
$= \Sigma_{i=0}^{n} (-1)^i \cdot (v_0, v_1, \ldots, \hat{v}_i, \ldots, v_n)$,
namely the sum of alternately +1 and -1 times **faces** of the original simplex formed by leaving out the i-th vertex

The boundary operator satisfies $im(\delta_{n+1}) \subset ker(\delta_n)$, so that boundary times boundary is zero: $\delta_n \circ \delta_{n+1} = 0$. (Note composition goes from right to left, in this case from $C_{n+1} \to C_n \to C_{n-1}$)

Therefore, for each n, you can form the quotient **homology class**
$H_n(X) = ker(\delta_n)/im(\delta_{n+1})$

A good choice for the boundary operator is the natural geometric boundary as calculated for each simplex. For example, δ of a 2-simplex (a triangle (a,b,c)) is a set of three 1-simplexes (line segments) $(b, c) - (a, c) + (a, b)$
Similarly, 0-simplexes are points (like (a), which is a point named "a", not a numeric value!), and a 1-simplex is an "edge" (a, b) with boundary as follows:
$\delta(a, b) = (b) - (a)$
Careful: you have to pay attention to sign!

Question 5: show that $\delta_n \circ \delta_{n+1}(a, b, c) = 0$ in this case (where n=1).
[Hint: the argument varies depending on n: δ_n is defined on an n-simplex.]

<---- SPOILER ALERT---->

$\delta_{n+1}(a, b, c) = (b, c) - (a, c) + (a, b)$

$\delta_n((b, c) - (a, c) + (a, b))$
$= \delta_n(b, c) - \delta_n(a, c) + \delta_n(a, b)$
$= ((c) - (b)) - ((c) - (a)) + ((b) - (a))$
$= (c) - (b) - (c) + (a) + (b) - (a)$
$= 0 \cdot (c) + 0 \cdot (b) + 0 \cdot (a) = 0$

In 'singular' homology you don't think of X as itself being made of simplexes, but rather you consider chains as composed of images of 'standard' n-dimensional hyper-tetrahedra mapped into X. This seems like a huge space, but when you take kernel 'divided' out by image, under reasonable circumstances, you get the same result. Advantages: this definition provides useful topological intuition, and you don't have to build X up as a simplicial complex.

If X is an n-dimensional manifold, you can think of k-dimensional submanifolds.

[Instead of integer multiples (our default!), you could use another 'ring of coefficients' like $\mathbb{R}$ or even $\mathbb{Z}/(2)$ to count multiples of simplices.]

--> Example from class: homology of S_1, D_2, used to demonstrate Brower's fixed-point theorem.

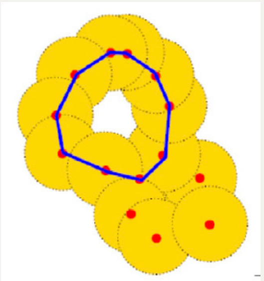
(We'll develop this more carefully in the second lecture.)

Topological data analysis

So what does any of this have to do with **TDA?**

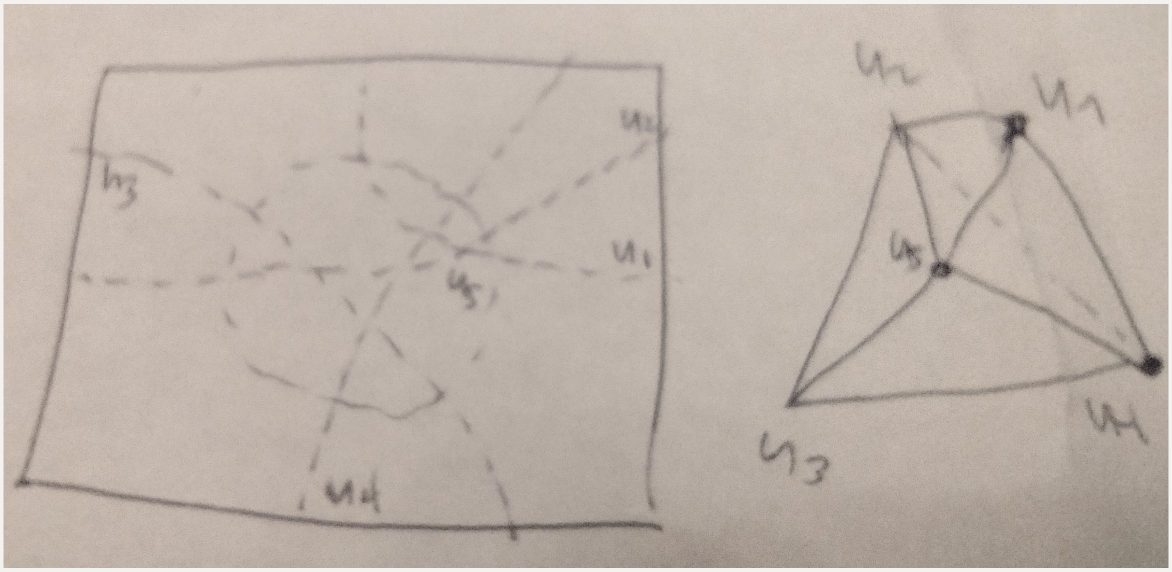
Answer: you can make simplicial complexes out of point spaces.

One way: given a set of points, take balls of a fixed radius around them. This gives you a topological space, which can be studied using AlgTop:



In this case, these balls are large enough to connect all the points, but they are also large enough to fill out (and thus obscure) the apparent lower ring formed by the points.

Another way to construct a topological space from a cover $\{U_i\}$ of a topological space X: the **nerve** (called the Cech complex):
Create an n-simplex for any set of n of the U_i that intersect



In this example, each U_i represents the area on the concave side of the curve's dotted line (with U_5 the interior of the circular region.) Note that $U_1 \cap U_3 = \emptyset$.
(Diagram corrected from Chazal & Michel.)

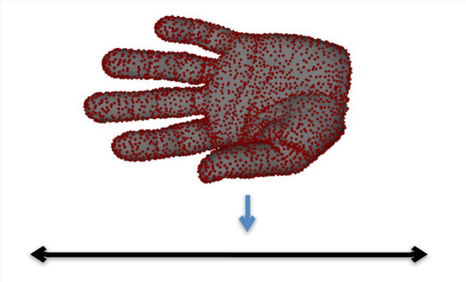
This can be a suprisingly faithful transformation of a space X from which a point cloud has been sampled. (But note the point cloud itself has a boring discrete topology.)

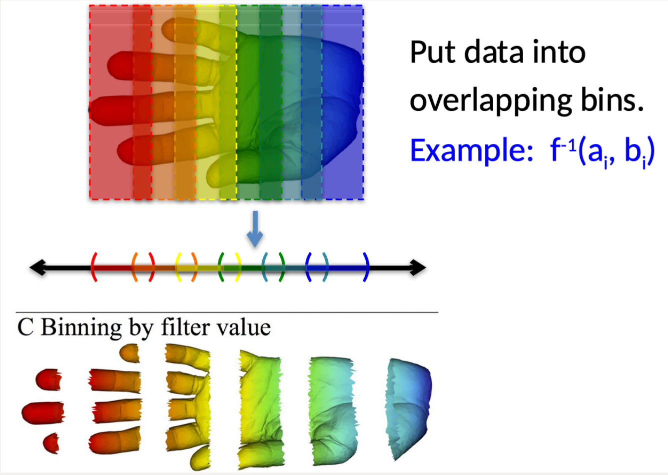
The **Nerve Theorem**: If $\{U_i\}$ is an open cover of X such that any nonempty intersection of subsets of $\{U_i\}$ is contractible. Then X and the nerve from this cover are homotopy equivalent.

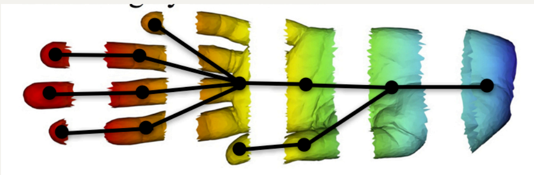
Lesson: by covering a set of points with open sets, you can do rigorous homology on the resulting topological space.
(The point set itself has the boring discrete topology.)

Mapper is one way to approach this, but it is useful as an exploratory tool only. It depends on choices of the cover and lenses, and is not guaranteed to pick up particular features.

Another mapper example: studying a point cloud sampled from the surface of a hand:







Question 6: What useful structure of the hand is captured and lost by the resulting graph

<---- SPOILER ALERT---->

- Captured: linear extensions ("flares") of fingers: these features are geometric, not topological
- Lost: the surface topology of the hand. Capturing this would likely take multiple lenses, and thus a much more computationally expensive setup.

Conclusion

You have seen a surprising connection between topology, algebraic topology, and data science.

This has helped grow the new subject of Topological Data Analysis, which continues to develop rapidly.

There is now a bit more science beyond this one technique, as we will explore more next week.

References

- A short overview with several pretty pictures, useful terms, and sample Python code you can run: https://www.quantmetry.com/blog/topological-data-analysis-with-mapper/
- From Frederic Chazal and Bertrand Michel:
  Slides from '16: https://hal.inria.fr/hal-01614384/file/Mapper.pdf
  A more advanced paper - most recently from '21: https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full
- Elementary Applied Topology ("EAT"), by Robert Ghrist (freely available at https://www2.math.upenn.edu/~ghrist/notes.html)
  A groundbreaking TDA text aimed at researchers - so not that "elementary"!) It discusses many interesting topics, but incompletely.

AlgTop texts:

- Greenberg & Harper
- Massey
- Munkres (friendly introduction)

There is also a good modern introduction to AlgTop by Burt Totaro of UCLA in *The Princeton Companion to Mathematics* (The only of his many papers I can't find online.)