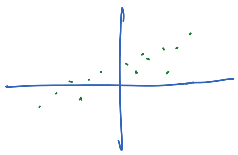
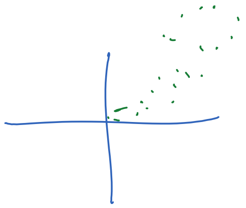
Topological Data Analysis

Thursday, January 11, 2018 4:13 PM

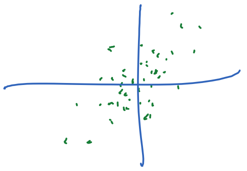
If you've done a lot of Data analysis, you know about



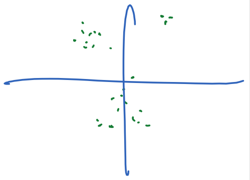
linear-ish data



exponential or power-law-ish data



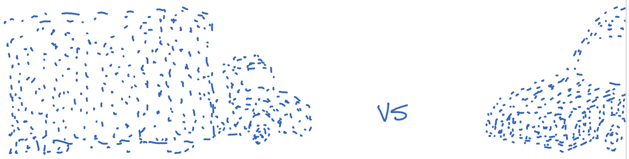
normal (Gaussian) data



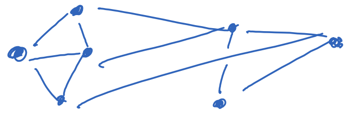
clustered data

But what about

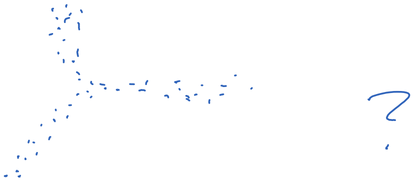
spatial data



network data



data with a shape



Topological data analysis (TDA) allows a fun

different way to look at this data.

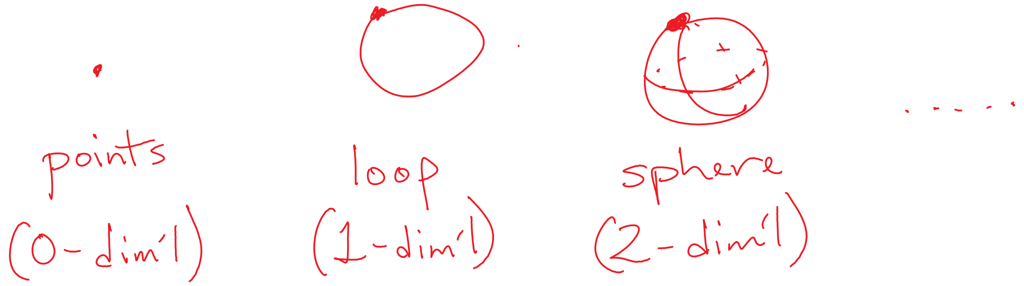
Used fruitfully on

- cancer data
- diabetes

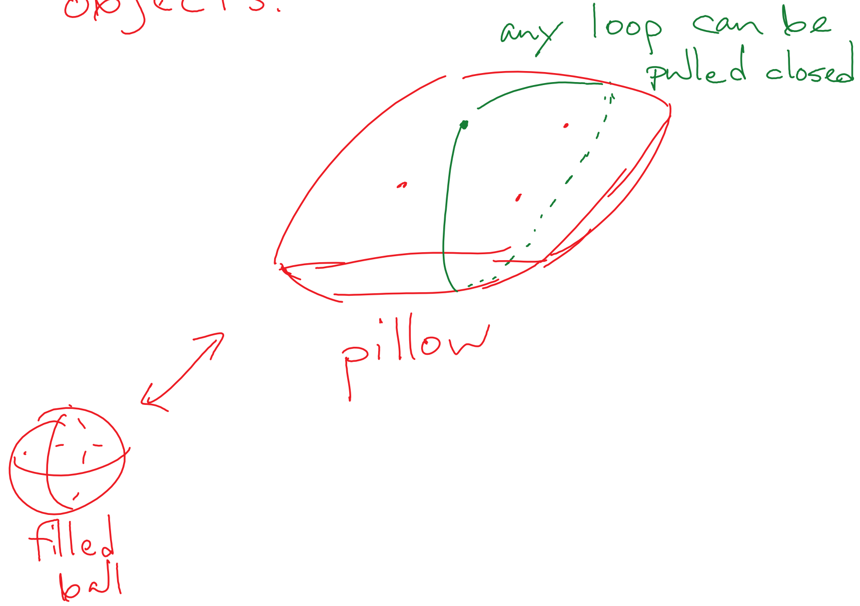
- ADHD
- tortuosity of blood vessels
- LiDAR data (ambulance vs Toyota T machine gun mount)

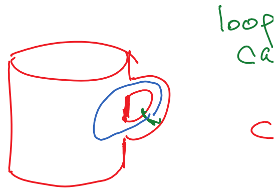
Topology crash course

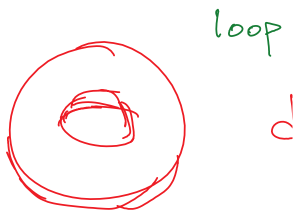
Pure math version: built on loops, spheres, etc.



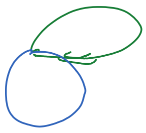
These help mathematicians differentiate between c objects.







blue loops
pull closed
on the s



The pillow has no way you

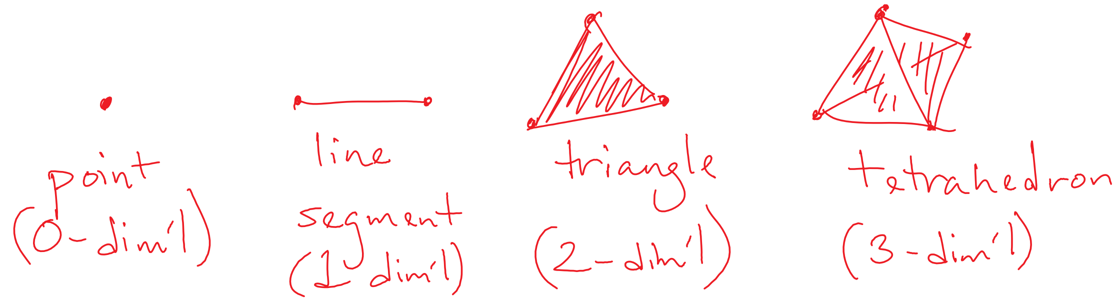
can make a loop on the surface that can't be pulled closed.

The coffee mug a
each have two
kinds of loops t
be pulled clos
surface.

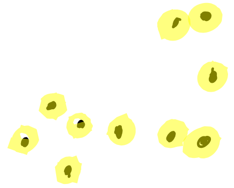
This pure math "homology" does not care about distances.

[homo ≈ same, so stru
objects are "the s

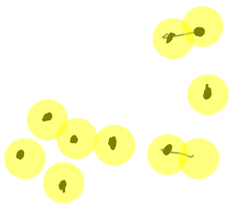
We need a metric (measure, length, distance) for data



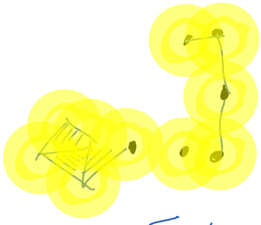
Vietoris-Rips complex



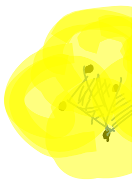
10 disconnected points



4 line segments have appeared



Lots of lines, and a few triangles.



Tetr
appe

Grow the radius of a ball around each point. Wh
connected?

Keeping track of how data points get connected to data points as you grow balls around them gives y topological "signature" for your data.

Persistent homology ~~barcode~~ or ~~birth-death~~ dia

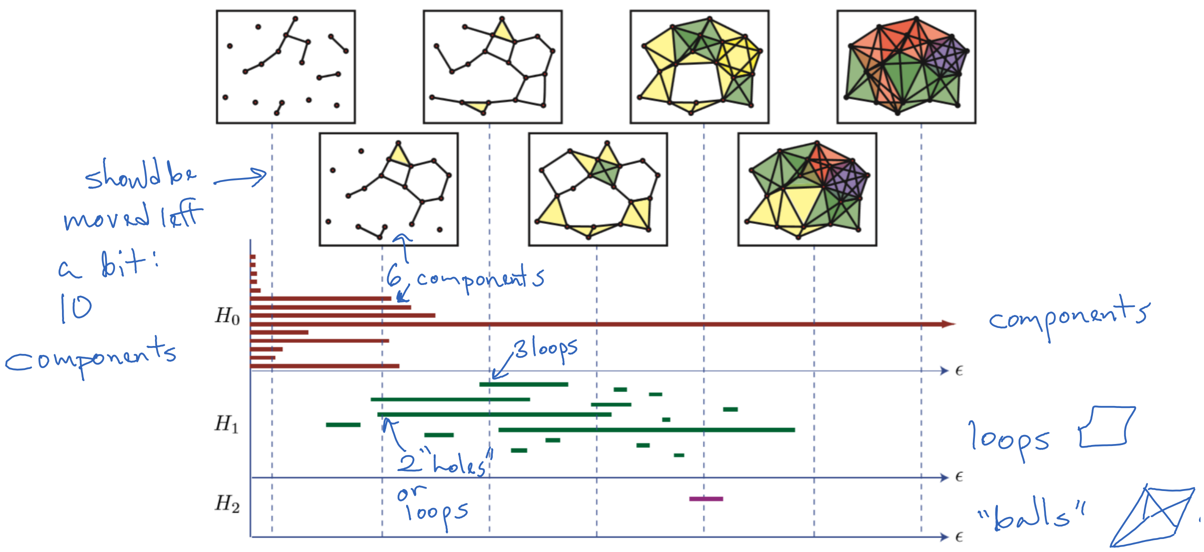
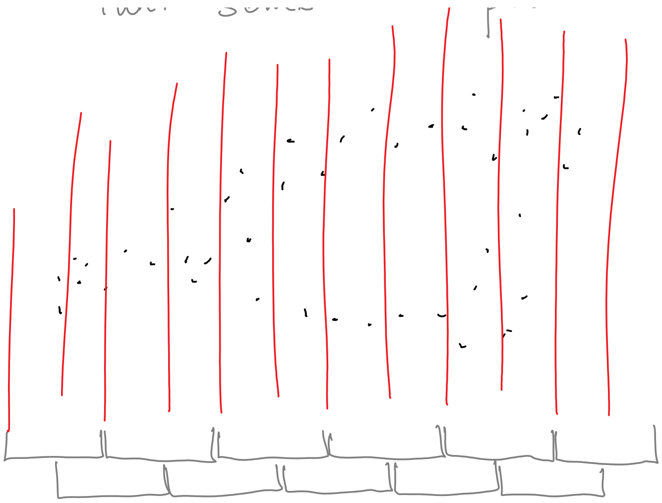


FIGURE 4. [bottom] An example of the barcodes for $H_*(\mathsf{R})$ in the example of Figure 3. [top] The rank of $H_k(\mathcal{R}_{\epsilon_i})$ equals the number of intervals in the barcode for $H_k(\mathsf{R})$ intersecting the (dashed) line $\epsilon = \epsilon_i$.

Screenshot from Rob Ghrist's "Barcodes: The persistent topology of data."

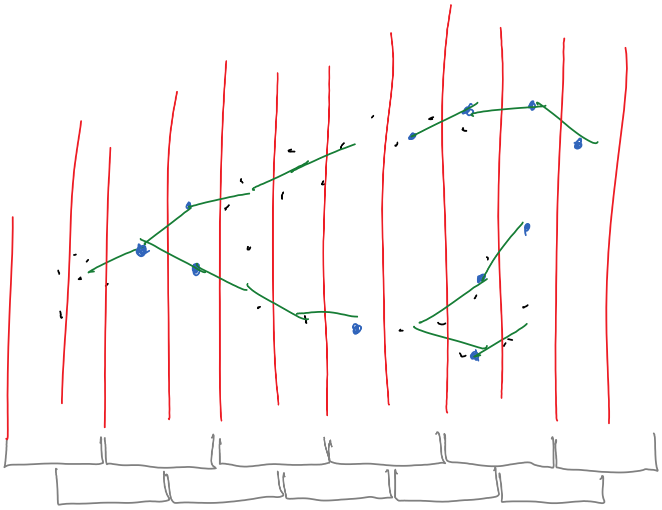
Workflow

- Prepare data if necessary (scaling, for instance)
- Choose a way to measure distance
- (• Project to 1 or 2 dimensions if you want to visua
- Choose a way to slice your data
- Tune some other, parameters



slice data like ||||| (as opposed to ☰ or ◎ or \\\\ ...)

with 50% overlap of bins



Once data is slic
clump in each s
and draw edges
there is overlap bet
clumps in overlappi

- Check the barcode to s
if topological features
persist or if you get
certain kind of signatu

- Check the graph you ge
against domain expertis
to see if it reveals
anything!