Topology for Data Science 2: Learning from the shape of data

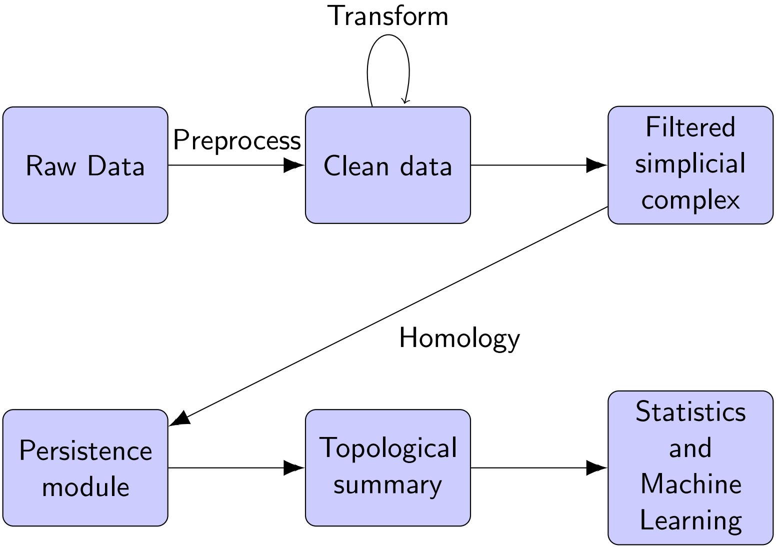
Peter Bubenik

University of Florida
Department of Mathematics,
peter.bubenik@ufl.edu
http://people.clas.ufl.edu/peterbubenik/

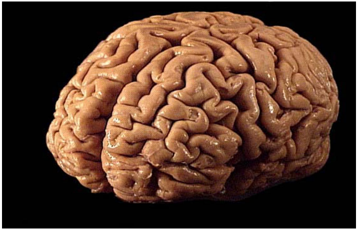
January 24, 2017

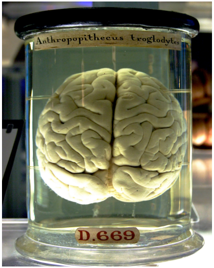
Tercera Escuela de Análisis Topológico de Datos
y Topología Estocástica
ABACUS, Estado de México

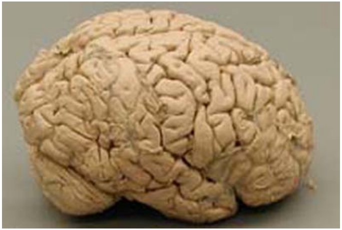
Topological Data Analysis



Brains





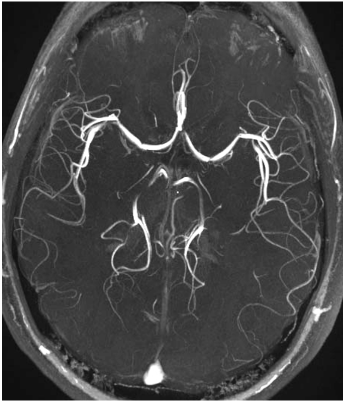


BRAINS!!!



Magnetic Resonance Imaging (MRI)

Brain arteries



Paul Bendich, J.S. Marron, Ezra Miller, Alex Pieloch, Sean Skwerer, Ann. Appl. Stat. **10** (2016) no. 1, 198–218 (presented here with changes by me)

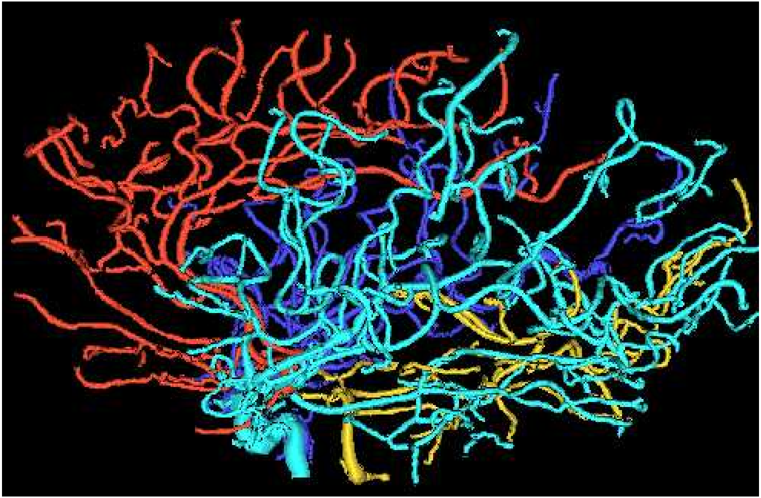
Brain arteries

Want to:

- understand normal changes with respect to age
- detect and locate pathology (tumors)
- predict stroke risk

Brain arterial data

Bullitt and Aylward (2002) MRA → Tubes



Mathematical viewpoint

Graph X with (x, y, z, r) for each vertex.

Mathematical viewpoint

Graph X with (x, y, z, r) for each vertex.

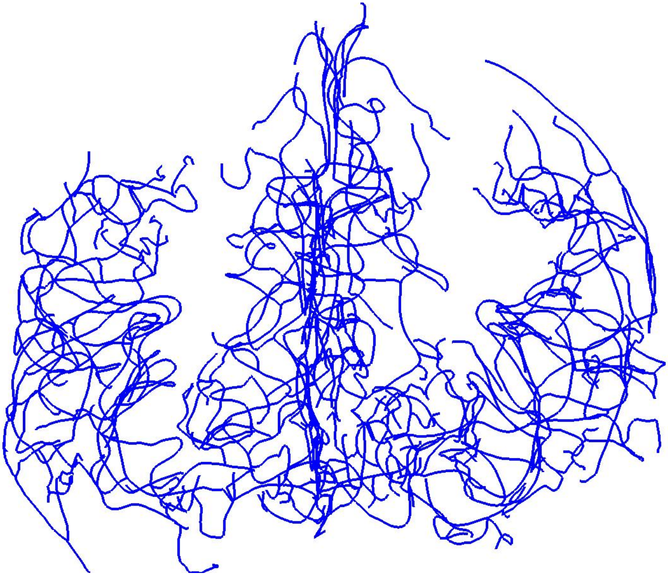
X_t is the full subgraph on the vertices with $z \leq t$.

$$\emptyset = X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_m = X$$

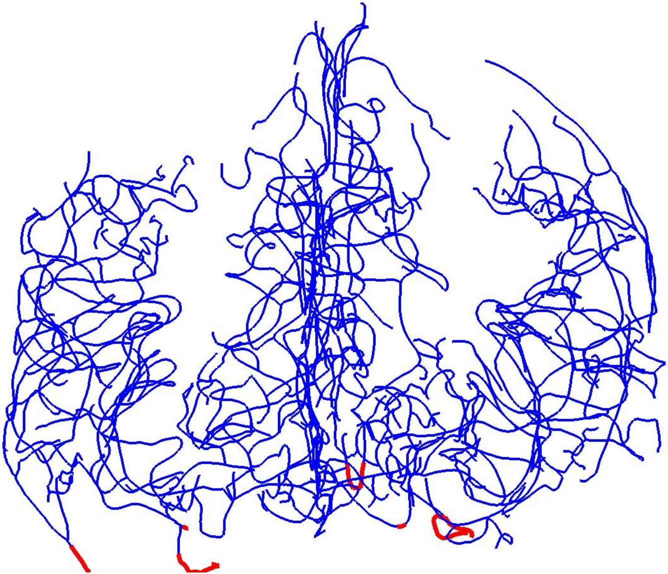
Apply H_0.

$$H_0(X_0) \to H_0(X_1) \to H_0(X_2) \to \cdots \to H_0(H_m)$$

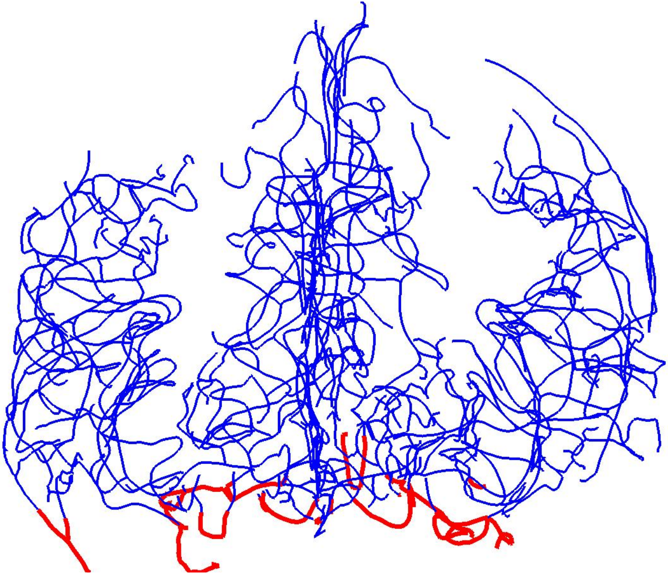
Filling the arteries – increasing sublevel sets



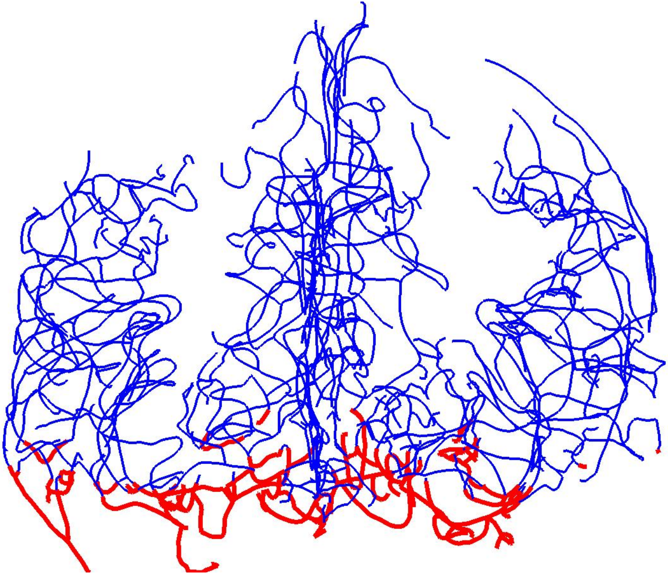
Filling the arteries – increasing sublevel sets



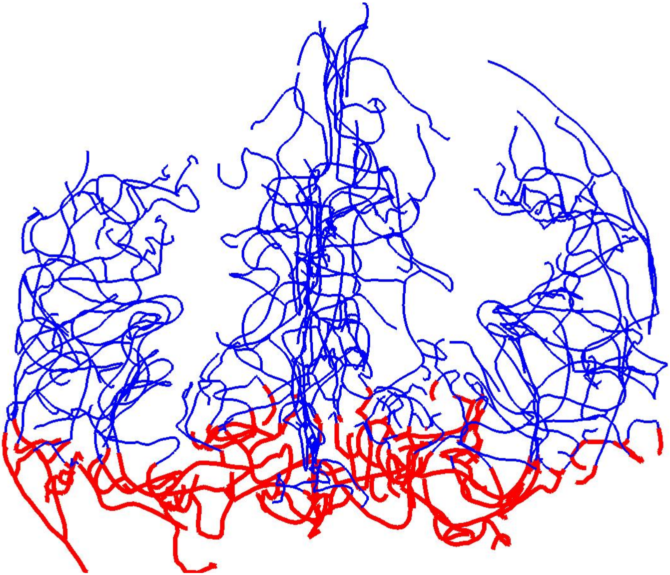
Filling the arteries – increasing sublevel sets



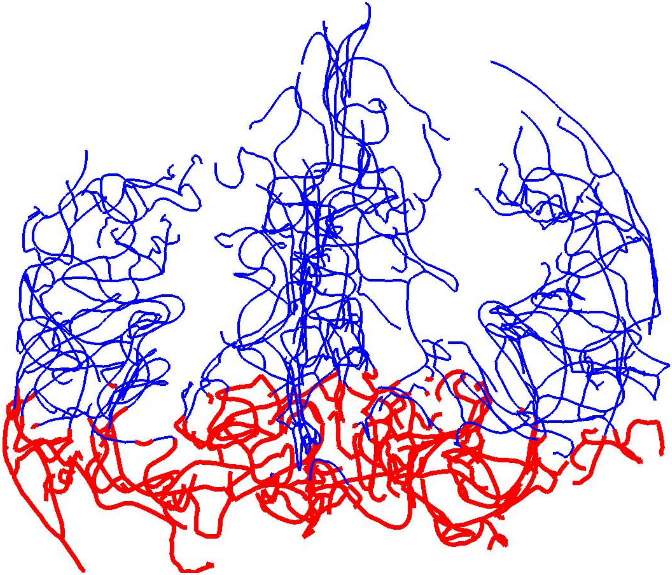
Filling the arteries – increasing sublevel sets



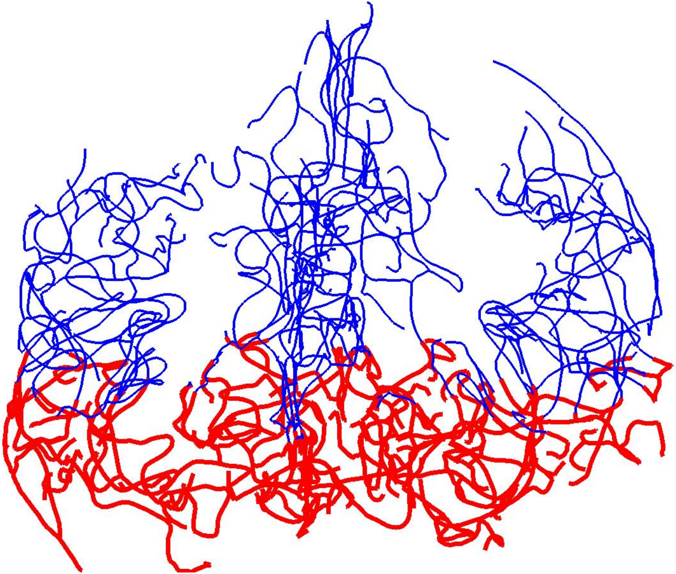
Filling the arteries – increasing sublevel sets



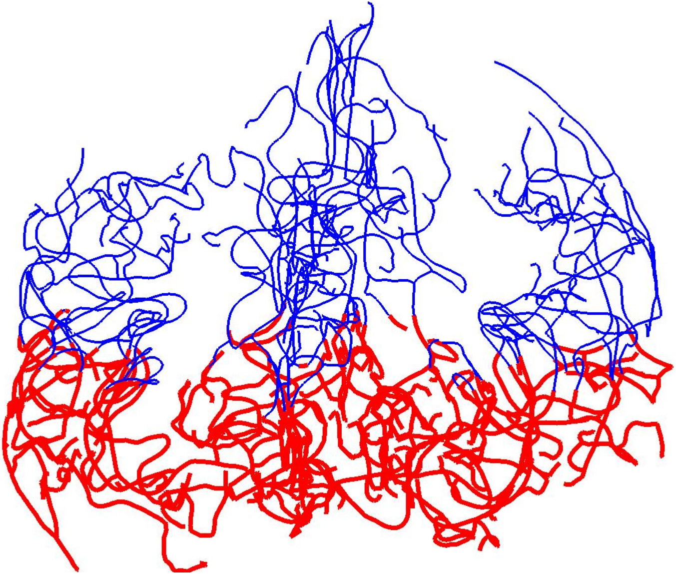
Filling the arteries – increasing sublevel sets



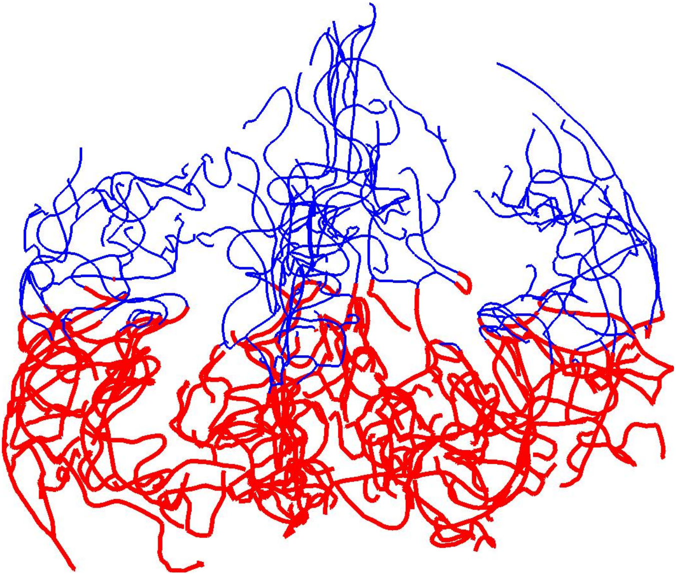
Filling the arteries – increasing sublevel sets



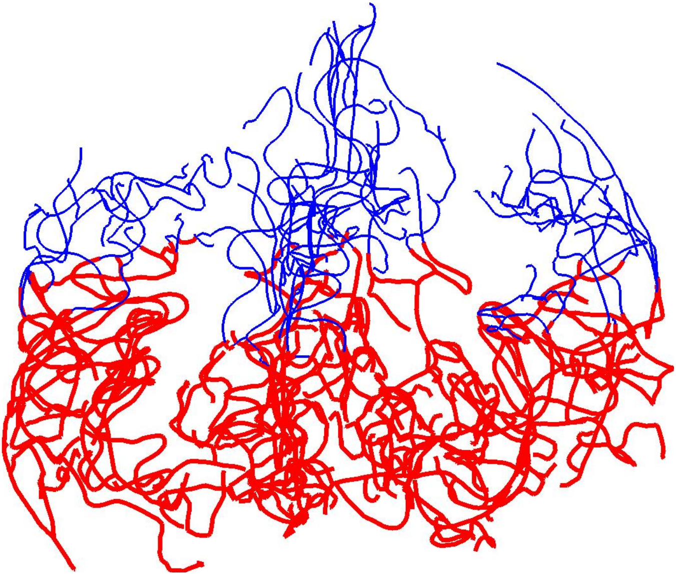
Filling the arteries – increasing sublevel sets



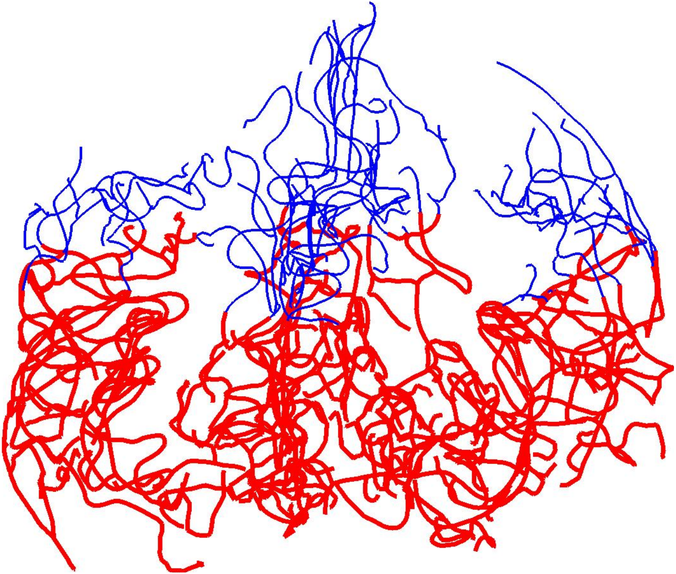
Filling the arteries – increasing sublevel sets



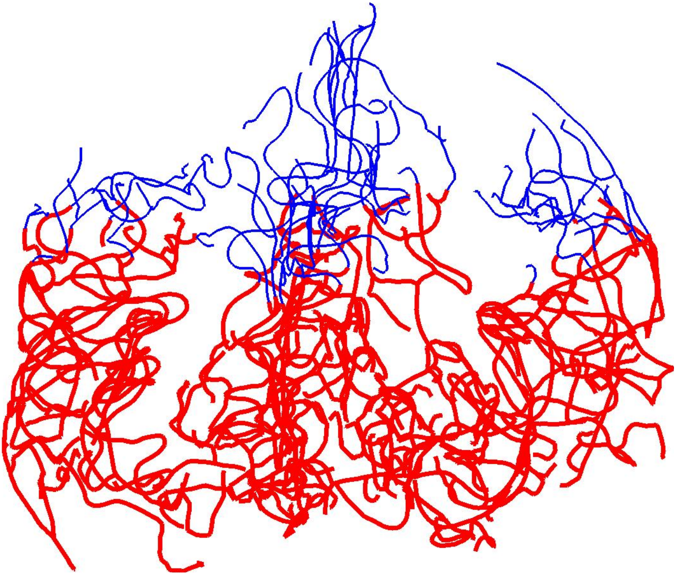
Filling the arteries – increasing sublevel sets



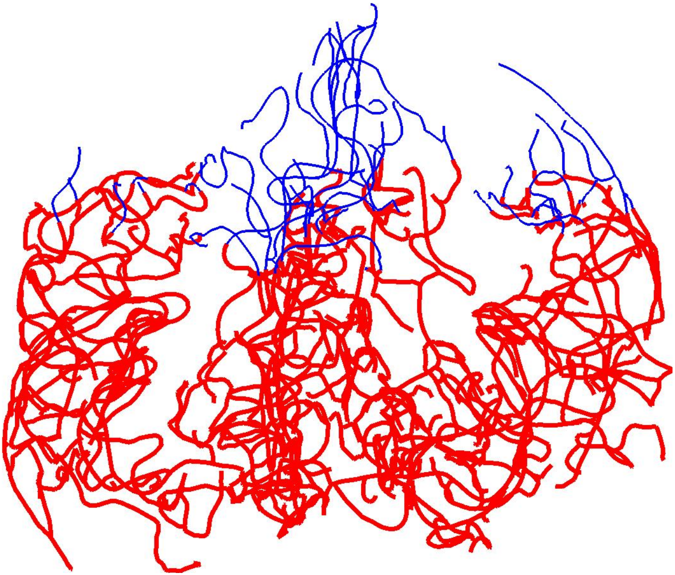
Filling the arteries – increasing sublevel sets



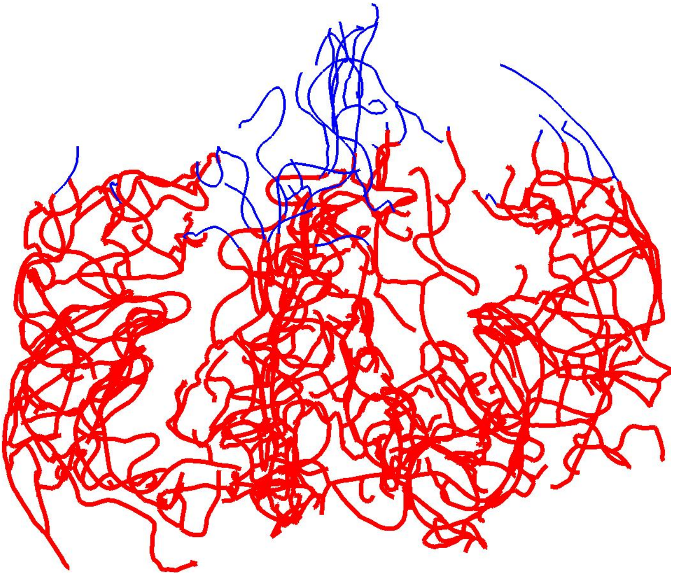
Filling the arteries – increasing sublevel sets



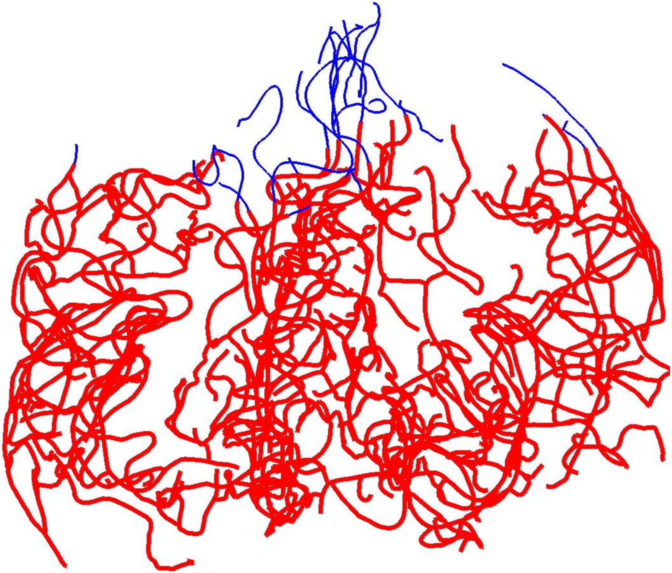
Filling the arteries – increasing sublevel sets



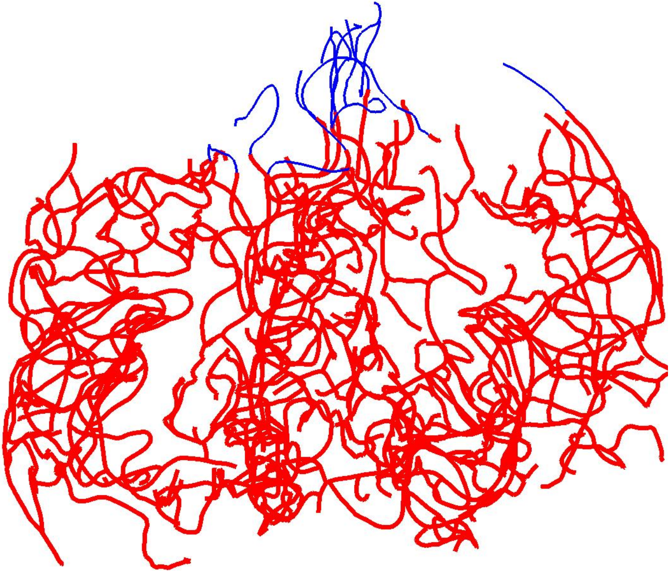
Filling the arteries – increasing sublevel sets



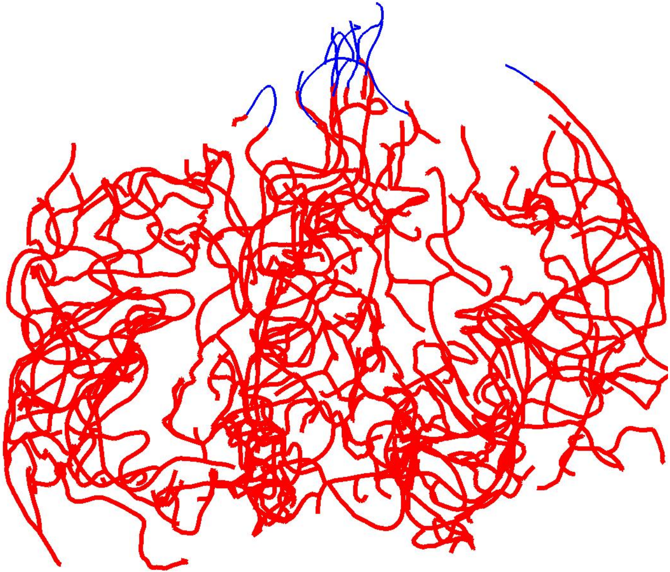
Filling the arteries – increasing sublevel sets



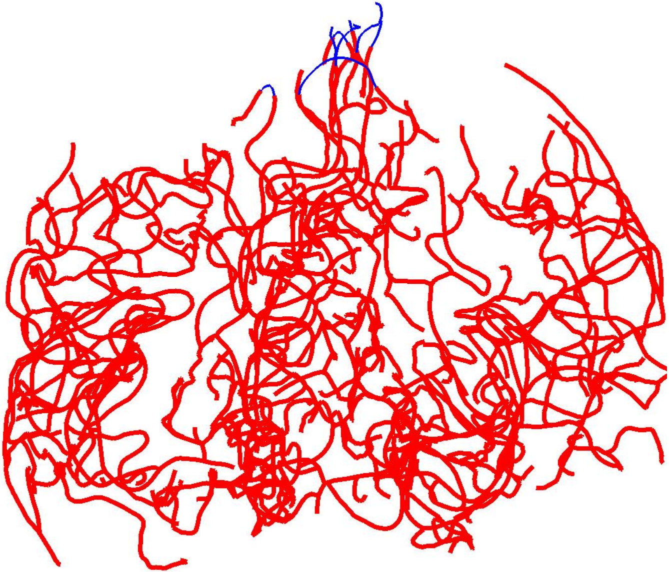
Filling the arteries – increasing sublevel sets



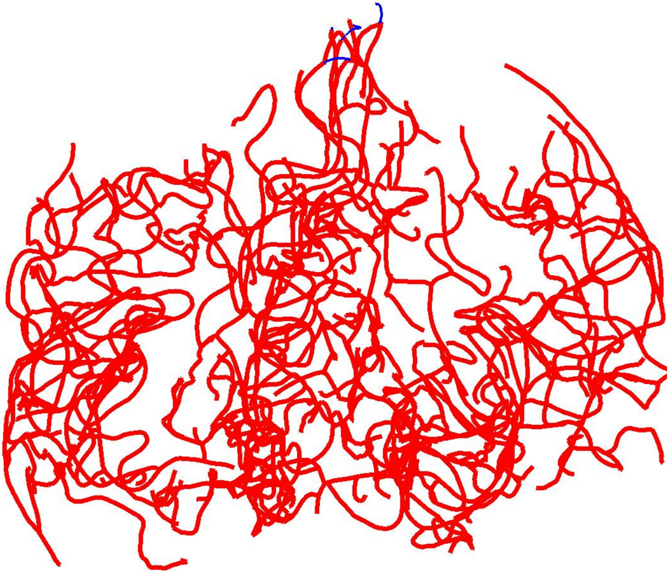
Filling the arteries – increasing sublevel sets



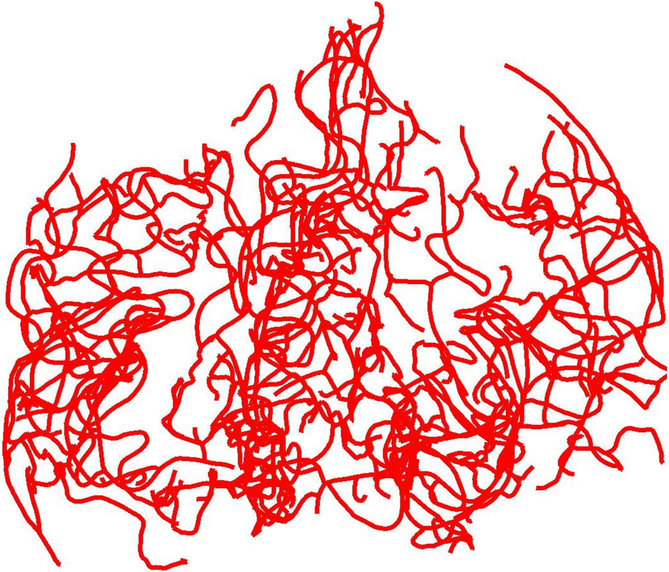
Filling the arteries – increasing sublevel sets



Filling the arteries – increasing sublevel sets



Filling the arteries – increasing sublevel sets



Mathematical encoding

We have an increasing sequence of simplicial complexes

$$X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_m$$

called a filtered simplicial complex.

Apply $H_k(-;\mathbb{F})$.

We get a sequence of vector spaces and linear maps

$$V_0 \to V_1 \to V_2 \to \cdots \to V_m$$

called a persistence module.

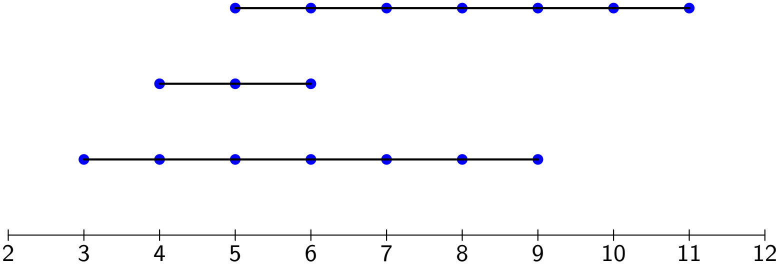
Summaries of persistence modules

$$V_0 \to V_1 \to V_2 \to V_3 \to V_4 \to V_5 \to V_6 \to V_7 \to \cdots \to V_m$$

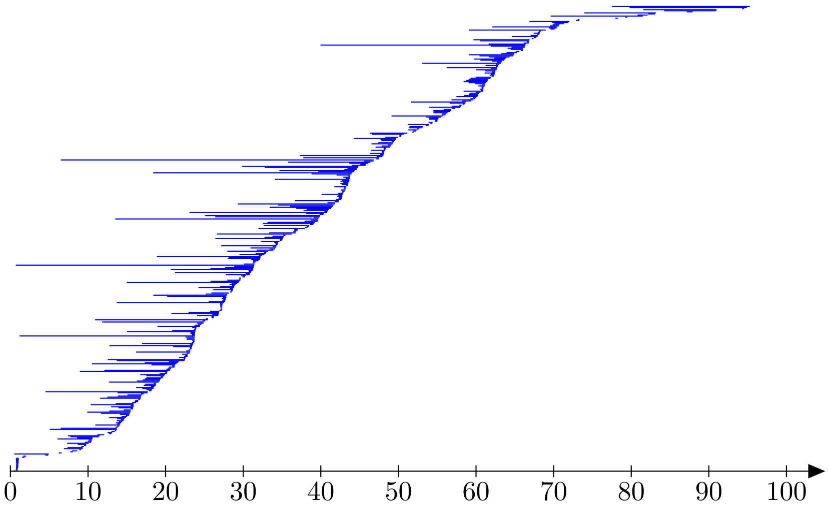
Fundamental Theorem of Persistent Homology

There exists a choice of bases for the vector spaces V_i such that each map is determined by a bipartite matching of basis vectors.

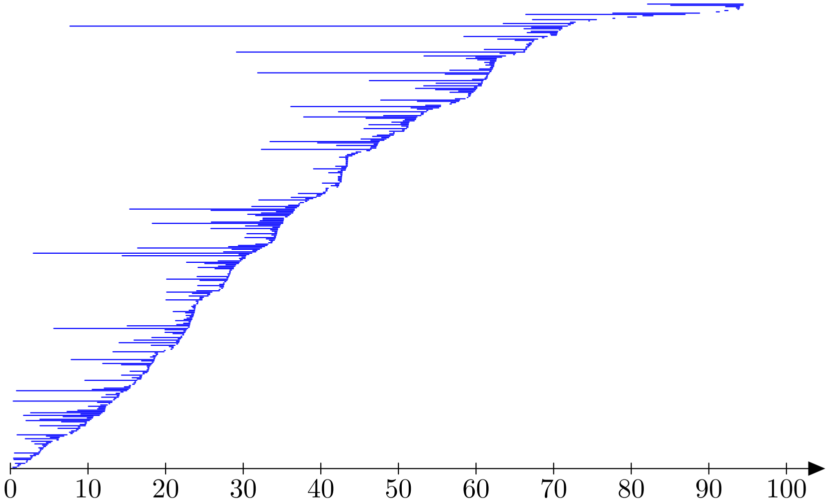
Get barcode.



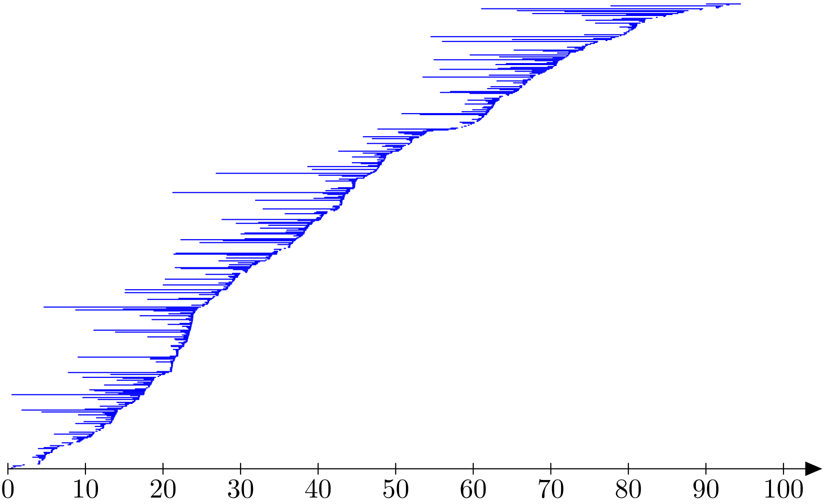
Brain artery barcodes



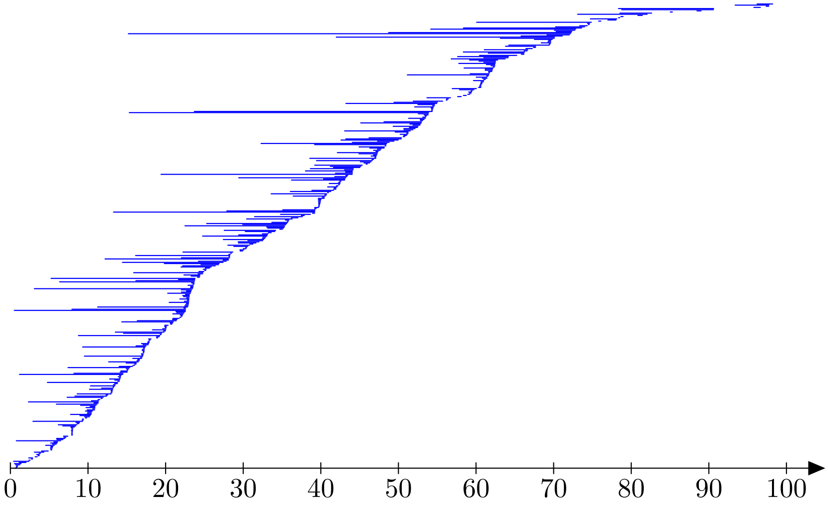
Brain artery barcodes



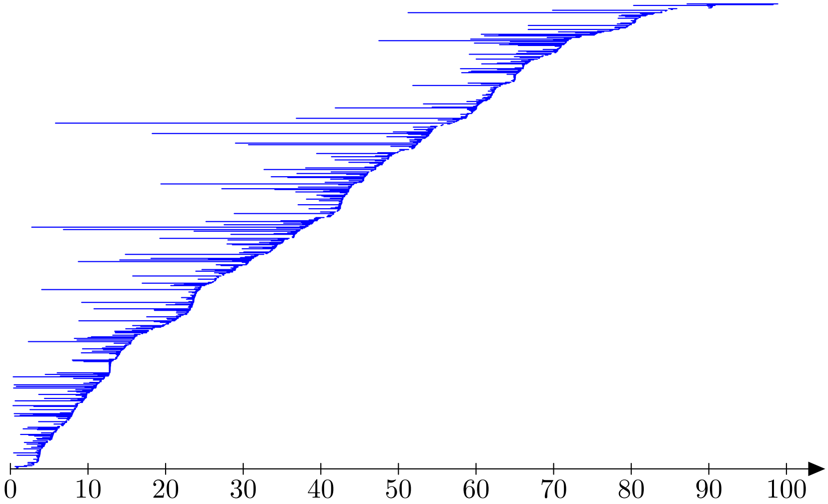
Brain artery barcodes



Brain artery barcodes



Brain artery barcodes



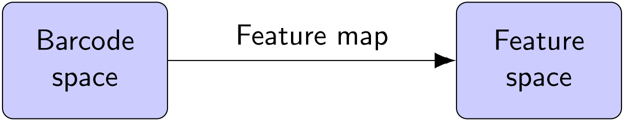
Challenges

We want to:

- calculate averages
- understand variances
- test hypotheses
- cluster and classify

We can only do a few of these using barcodes.

Making life easier



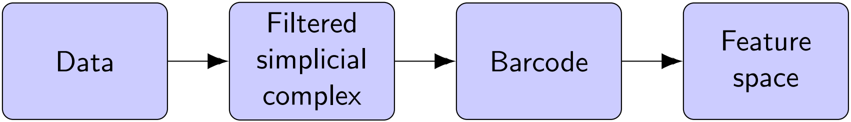
The feature space is a Hilbert space.

What is great about a Hilbert space?

- it is a vector space (easy to measure distances, averages)
- it has an inner product (easy to measure angles)
- it is complete (good for studying convergence)

Feature maps and kernels

Feature map $\Phi : \mathcal{X} \to \mathcal{H}$

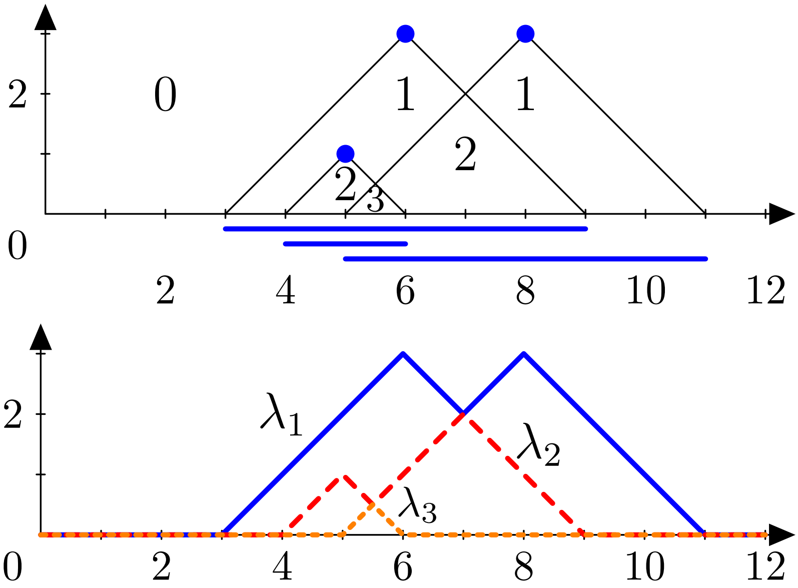


Kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\quad K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$.

With a kernel, we can

- measure distance $d(x, y) = K(x - y, x - y)^{1/2}$
- measure angles: $\angle(x, y) = \arccos \dfrac{K(x, y)}{K(x, x)^{1/2} K(y, y)^{1/2}}$

Constructing a persistence landscape



Constructing a persistence landscape