Statistical topological data analysis using persistence landscapes applied to brain arteries

CANSSI–SAMSI Workshop: Geometric Topological

and Graphical Model Methods in Statistics

Peter Bubenik

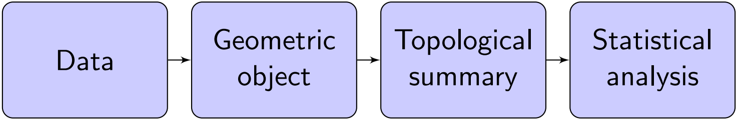
Department of Mathematics
Cleveland State University
p.bubenik@csuohio.edu
http://academic.csuohio.edu/bubenik_p/

May 23, 2014

funded by AFOSR

Statistical topological data analysis

The plan:



Brain arteries

Joint work with Ezra Miller (Duke/SAMSI), J.S. Marron (UNC-CH), Paul Bendich (Duke) and Sean Skwerer (UNC-CH).

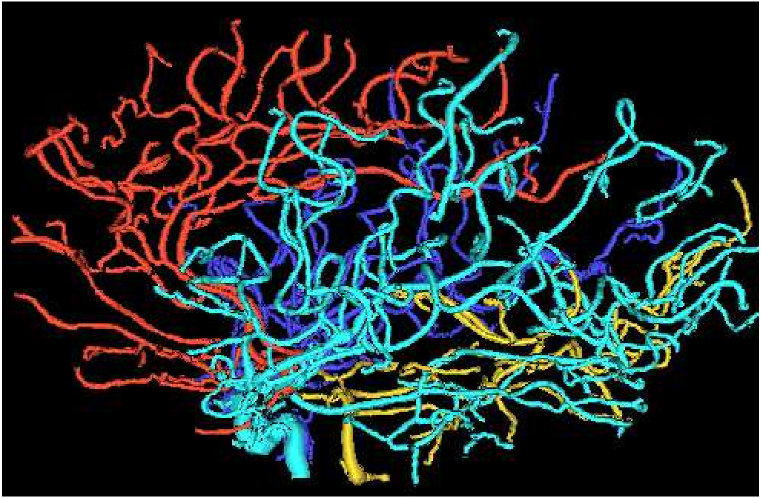
Brain arteries

Goal: Analyze the shape of brain arteries in order to

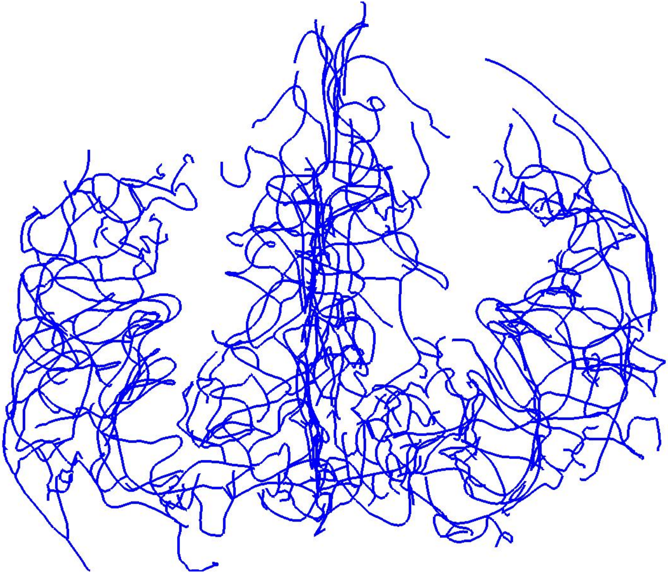
- understand normal changes with respect to age
- detect and locate pathology (tumors)
- predict stroke risk

The data

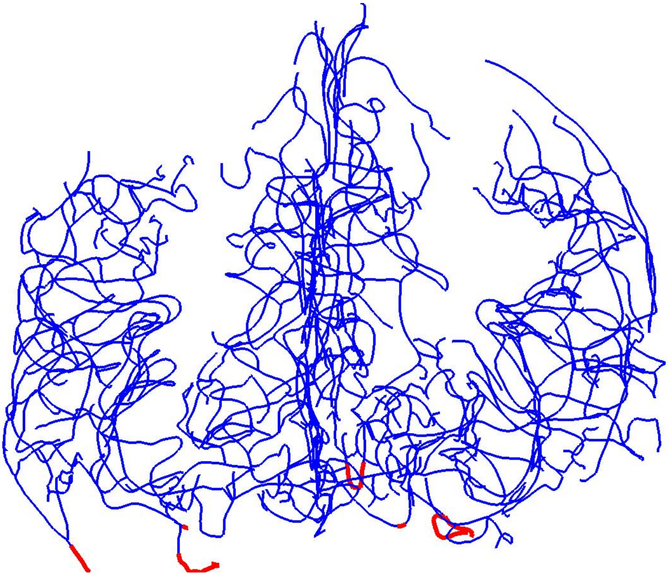
Bullitt and Aylward (2002) MRA $\rightarrow$ Tubes



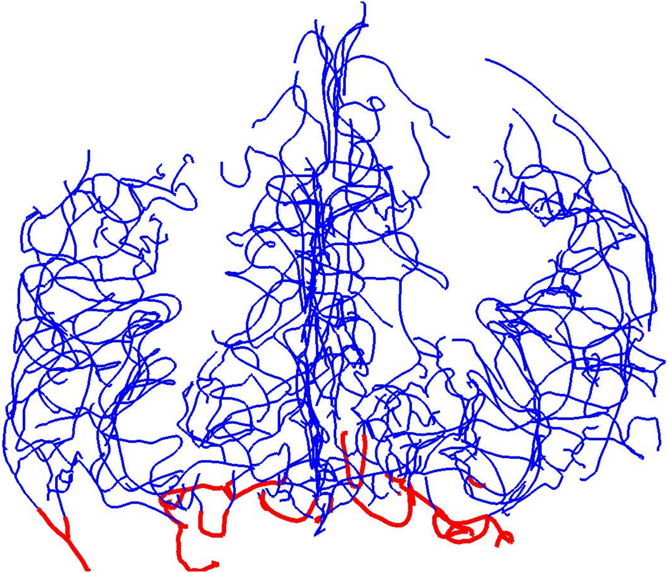
Filling the arteries – increasing sublevel sets



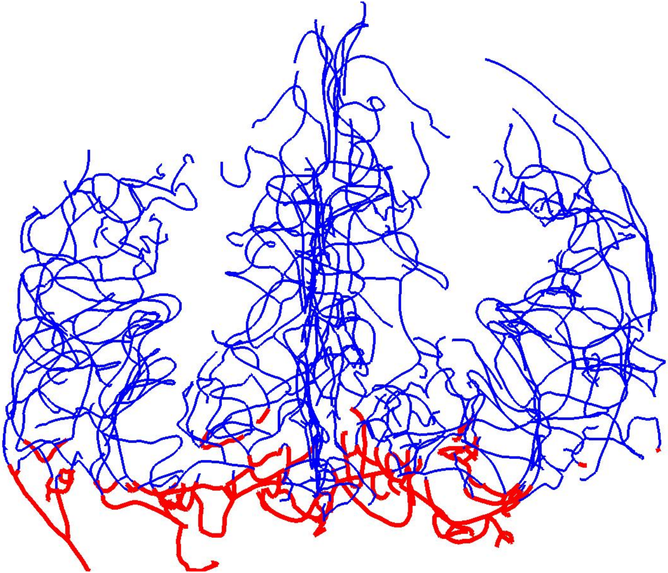
Filling the arteries – increasing sublevel sets



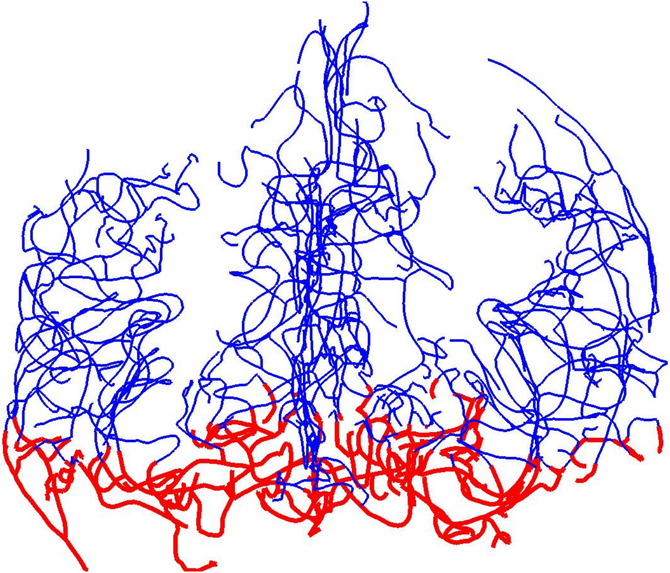
Filling the arteries – increasing sublevel sets



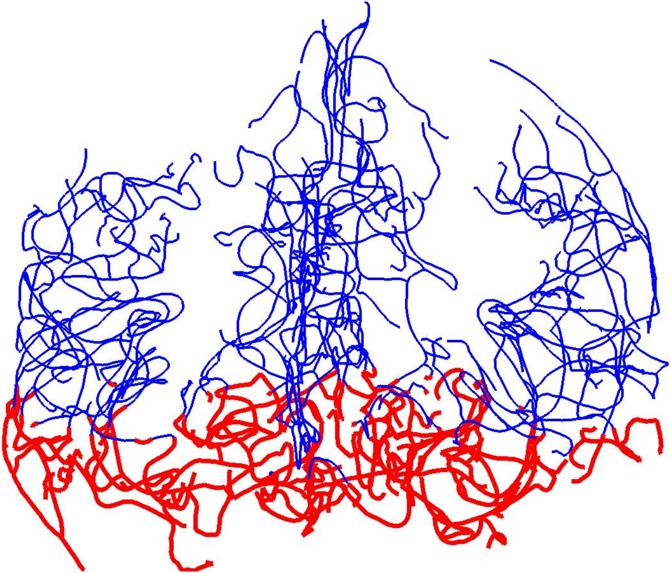
Filling the arteries – increasing sublevel sets



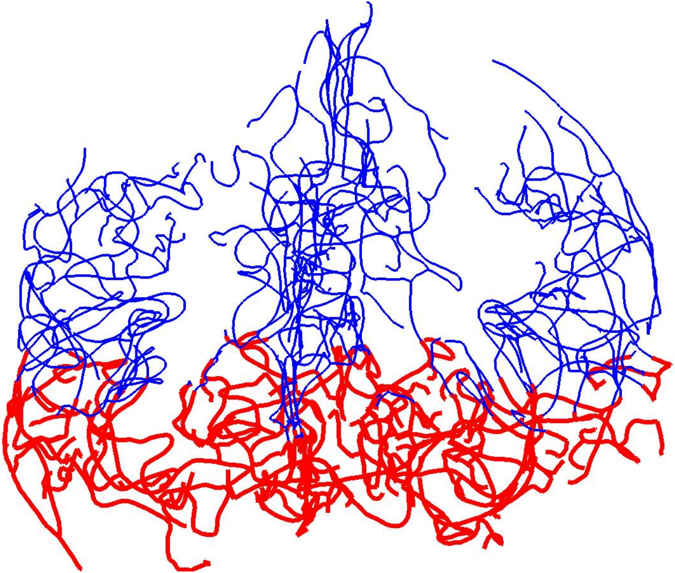
Filling the arteries – increasing sublevel sets



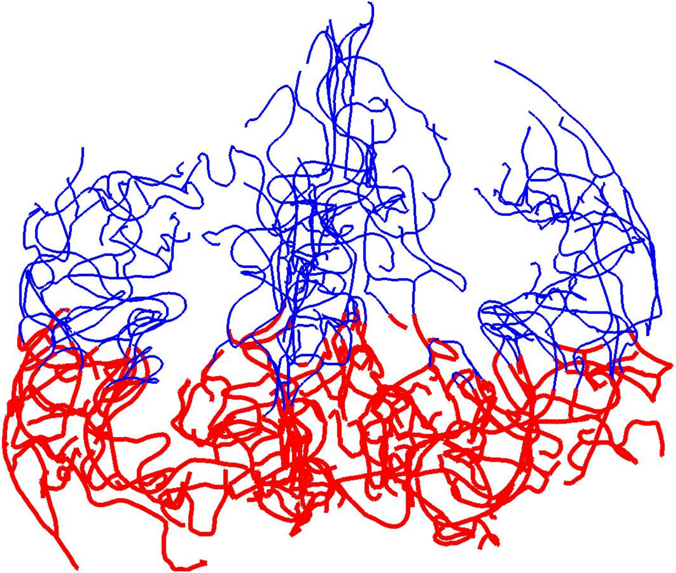
Filling the arteries – increasing sublevel sets



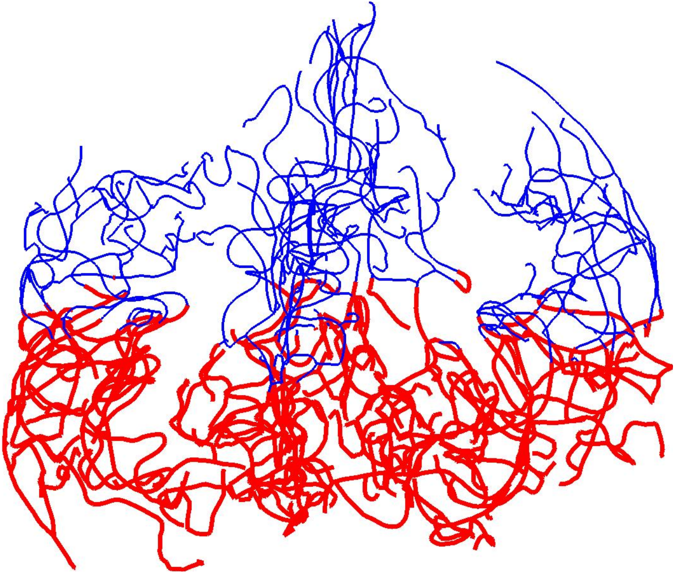
Filling the arteries – increasing sublevel sets



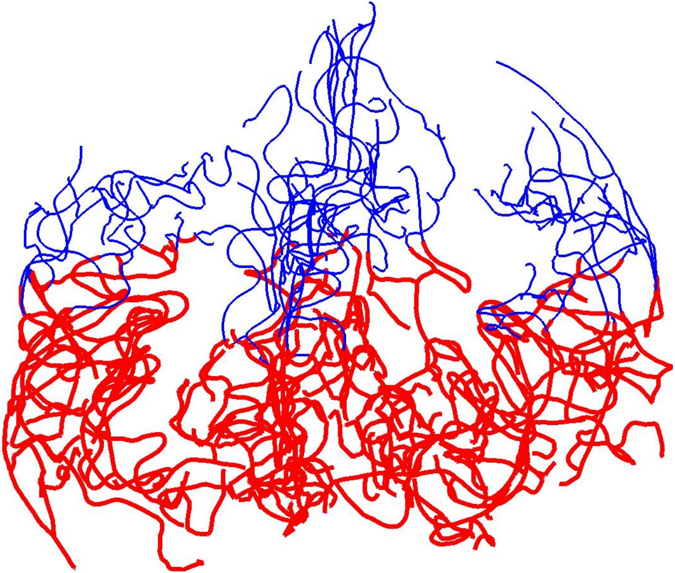
Filling the arteries – increasing sublevel sets



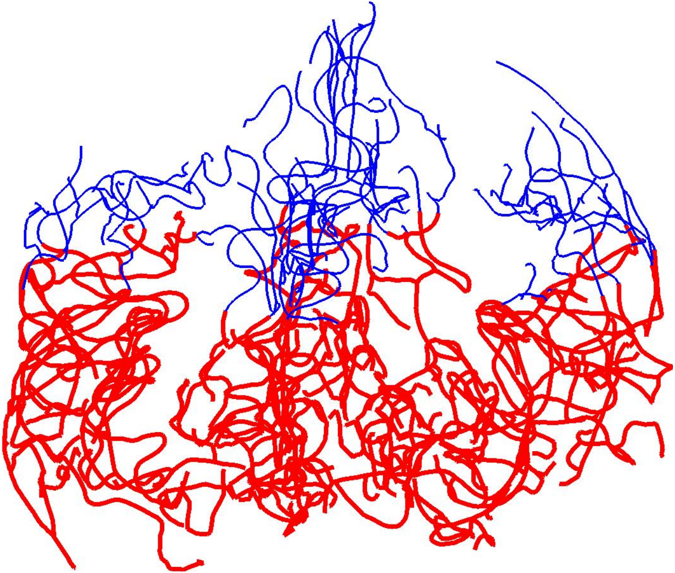
Filling the arteries – increasing sublevel sets



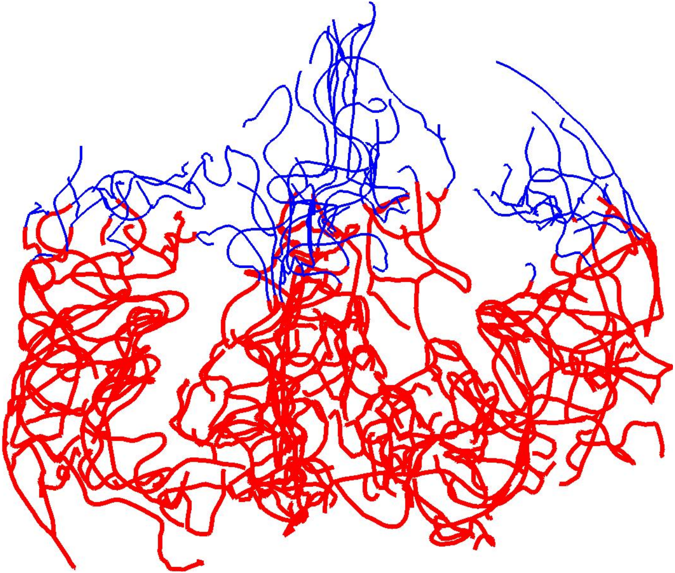
Filling the arteries – increasing sublevel sets



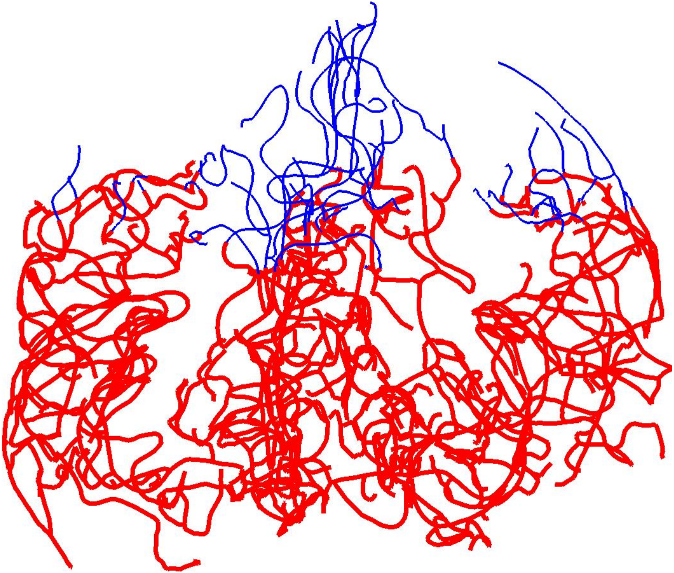
Filling the arteries – increasing sublevel sets



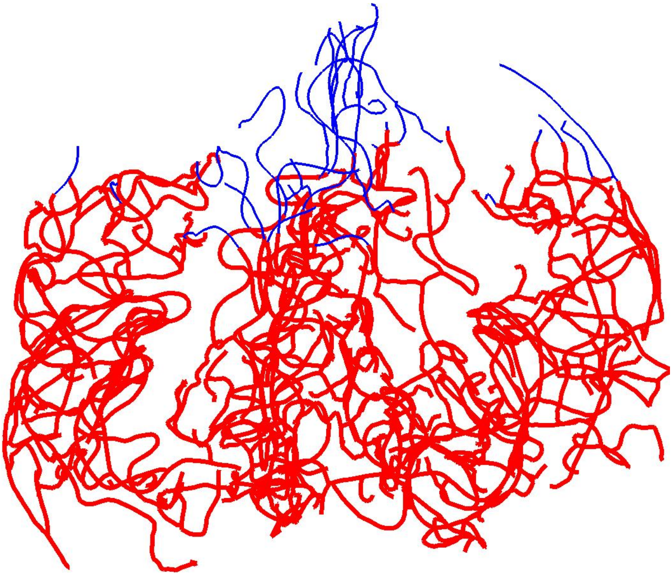
Filling the arteries – increasing sublevel sets



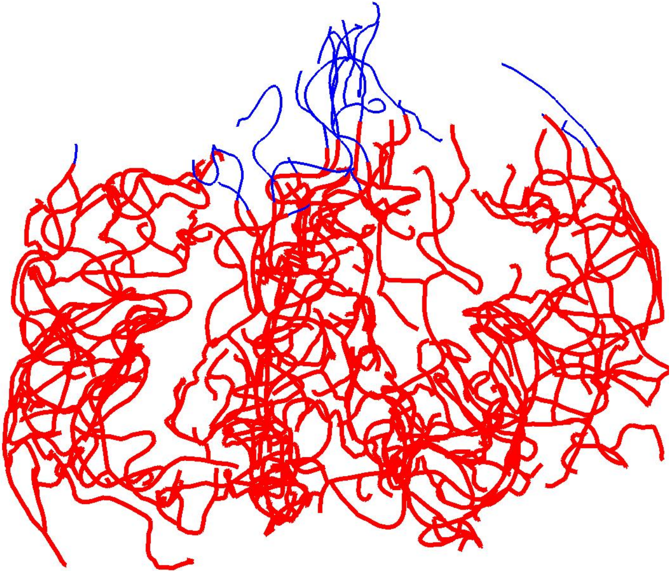
Filling the arteries – increasing sublevel sets



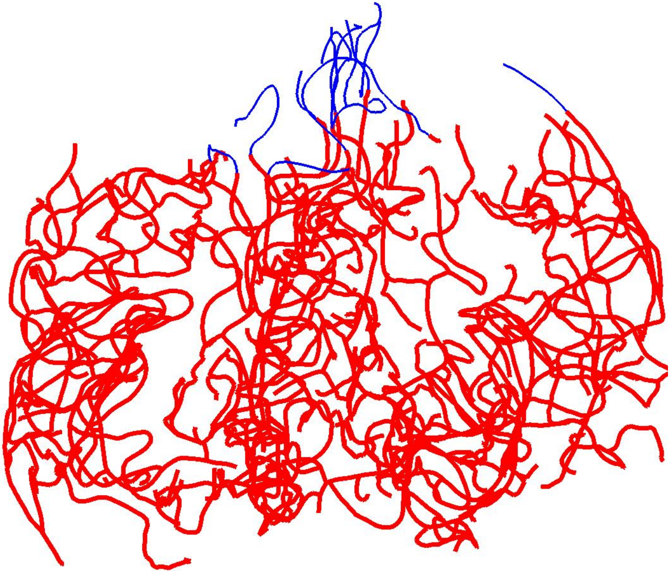
Filling the arteries – increasing sublevel sets



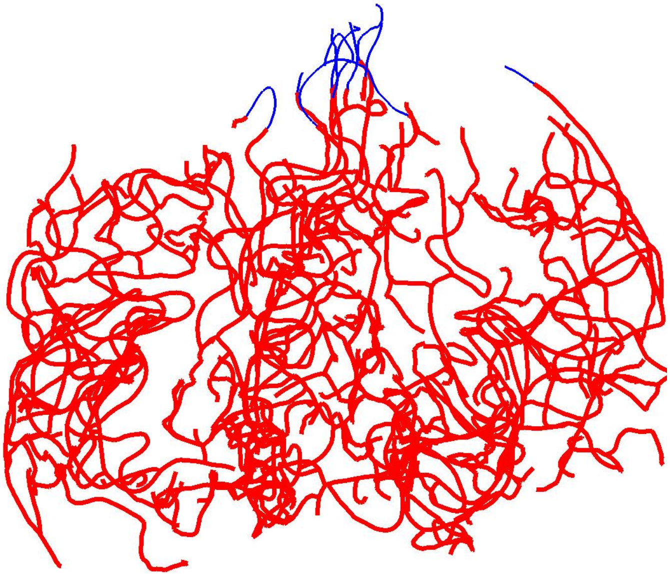
Filling the arteries – increasing sublevel sets



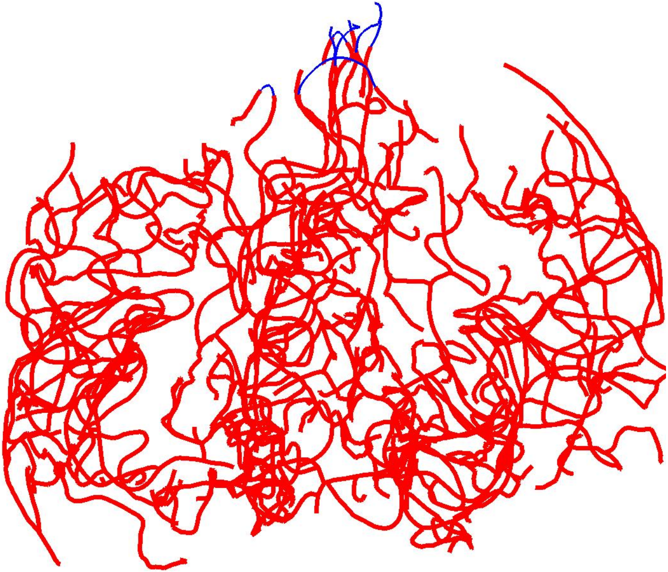
Filling the arteries – increasing sublevel sets



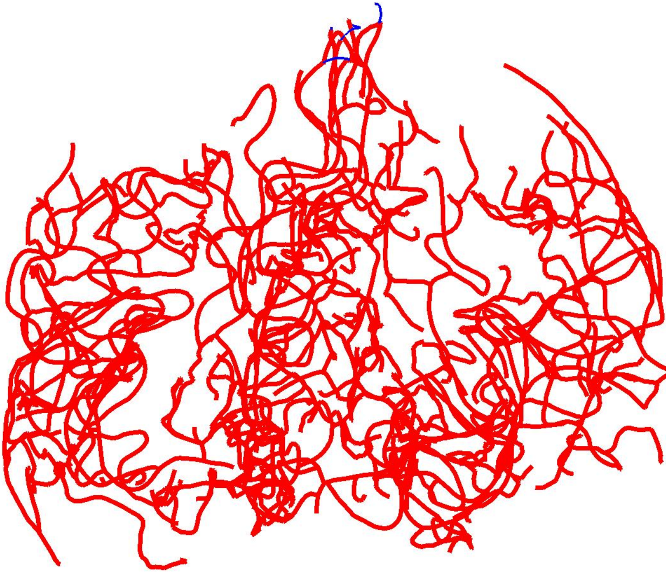
Filling the arteries – increasing sublevel sets



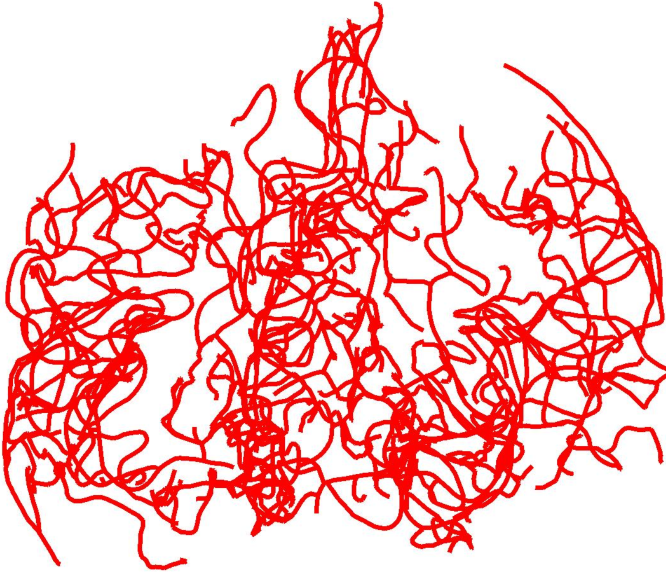
Filling the arteries – increasing sublevel sets



Filling the arteries – increasing sublevel sets



Filling the arteries – increasing sublevel sets



Mathematical viewpoint

Let X be a graph representing the brain arteries of one subject:

- vertices with (x, y, z, r) coordinates
- edges connecting adjacent vertices

Mathematical viewpoint

Let X be a graph representing the brain arteries of one subject:

- vertices with (x, y, z, r) coordinates
- edges connecting adjacent vertices

Let X_t denotes the full subgraph on the vertices with z coordinate at most t.

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

Take homology in degree 0.

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

More general setup

For each t, have
- a simplicial complex X_t
- a vector space $H(X_t)$

For $t \leq t'$, have
- an inclusion $X_t \subseteq X_{t'}$
- a linear map $H(X_t) \to H(X_{t'})$

Persistent homology is the image of this map.

This set of vector spaces and linear maps is called a persistence module.

We want a summary of the persistence module that is amenable to statistical analysis.

Persistence landscape

Recall that the persistence module consisted of linear maps

$$H(X_t) \to H(X_{t'}), \text{ for } t \leq t'.$$

For $k = 1, 2, 3, \ldots$, define $\lambda_k : \mathbb{R} \to \mathbb{R}$ by

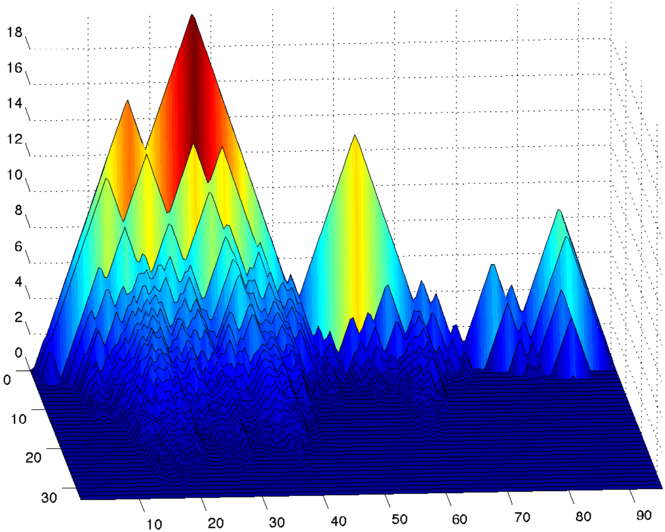
$$\lambda_k(t) = \max(\ h \ \mid \ \text{rank}(H(X_{t-h}) \to H(X_{t+h}) \geq k\ )$$

We can combine these to get one function

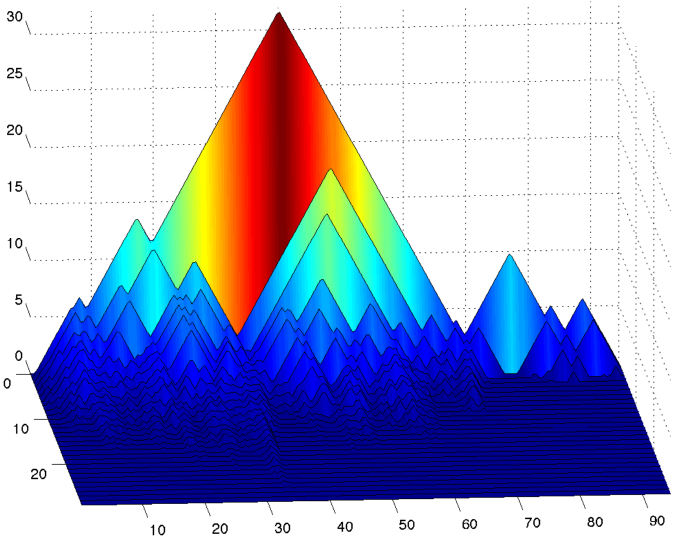
$$\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R},$$

where $\lambda(k, t) = \lambda_k(t)$.

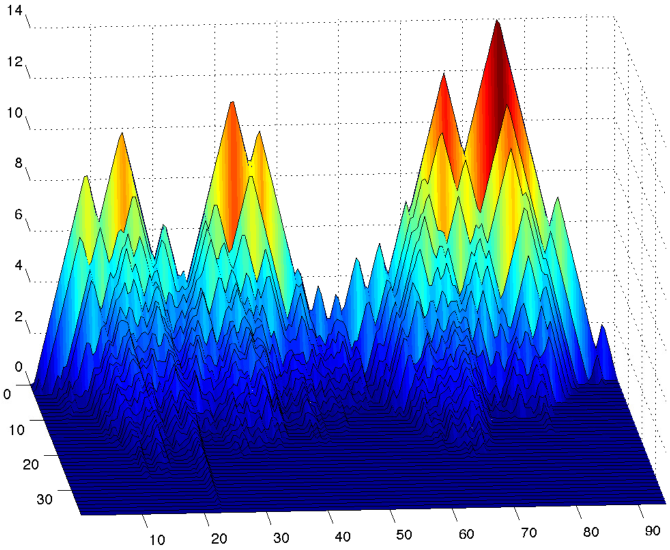
Persistence landscape examples



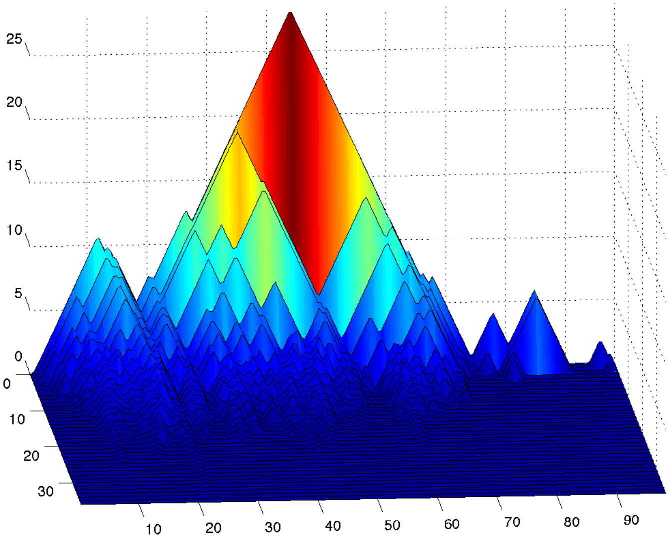
Persistence landscape examples



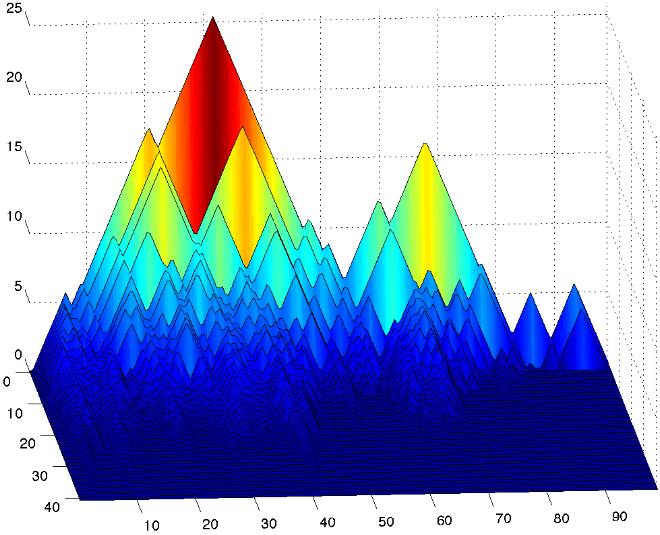
Persistence landscape examples



Persistence landscape examples



Persistence landscape examples



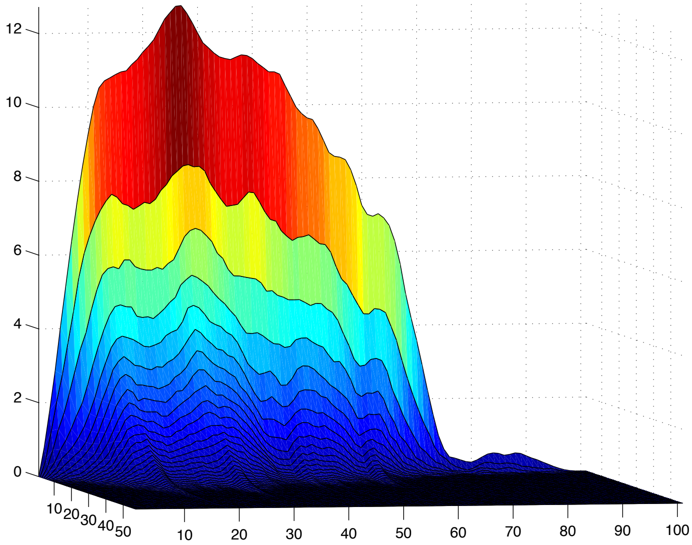
Mean landscapes

Persistence landscapes, $\lambda^{(1)}, \ldots, \lambda^{(n)}$, have mean, $\overline{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \lambda^{(i)}$.

That is,

$$\overline{\lambda}_k(t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_k^{(i)}(t)$$

Mean landscape for brain arteries



Summary space

Let $1 \leq p < \infty$. Then $\|\lambda\|_p = \left(\sum_k \int {\lambda_k}^p \right)^{\frac{1}{p}}$.

We assume $\|\lambda\| := \|\lambda\|_p < \infty$. That is, $\lambda \in L^p(\mathbb{N} \times \mathbb{R})$.

So λ is a random variable with values in a Banach space.

Asymptotics

$\lambda \in L^p(\mathbb{N} \times \mathbb{R})$, $\|\lambda\|$ is a real random variable.

If $E\|\lambda\| < \infty$ then there exists $E(\lambda) \in L^p(\mathbb{N} \times \mathbb{R})$ such that $E(f(\lambda)) = f(E(\lambda))$ for all continuous linear functionals f.

Theorem (Strong Law of Large Numbers (SLLN))

$\overline{\lambda}^{(n)} \to E(\lambda)$ *almost surely if and only if* $E\|\lambda\| < \infty$.

Theorem (Central Limit Theorem (CLT))

Assume $p \geq 2$. *If* $E\|\lambda\| < \infty$ *and* $E(\|\lambda\|^2) < \infty$ *then* $\sqrt{n}[\overline{\lambda}^{(n)} - E(\lambda)]$ *converges weakly to a Gaussian random variable with the same covariance structure as* λ.

Weighted norms

Recall that $\|\lambda\|_p = \left(\sum_k \int \lambda_k{}^p \right)^{\frac{1}{p}}$.

Fix $i \leq j$. Define $\|\lambda\|_{p,i,j} = \left(\sum_{k=i}^{j} \int \lambda_k{}^p \right)^{\frac{1}{p}}$.

The previous SLLN and CLT also apply to this weighted norm.

Correlation with age

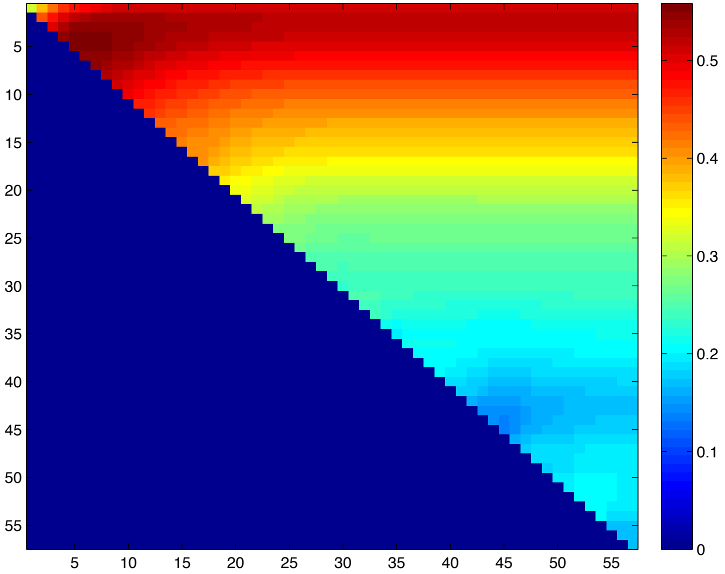
Pearson's correlation coefficient of
age with statistics derived from the brain arteries

Previous study without topology:
Dan Shen et al (2014) $r = 0.25$

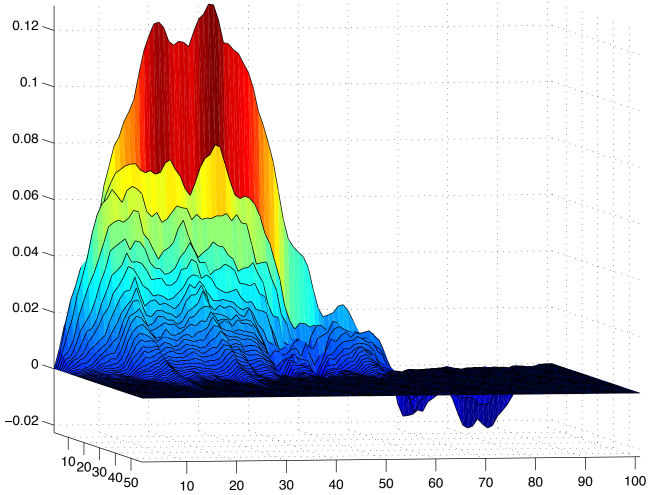
Using persistence landscape:

topological statistic	r
$\|\lambda\|_1$	0.5077
$\|\lambda\|_{1,2,57}$	0.5214
$\|\lambda\|_{1,5,5}$	0.5582

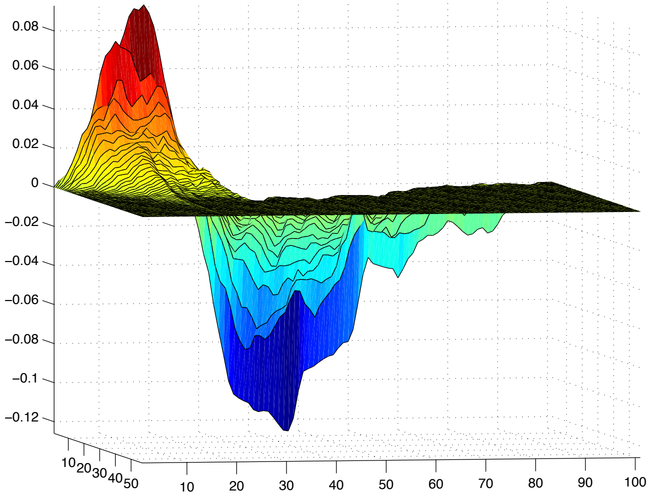
Correlation of age with $\|\lambda\|_{1,i,j}$



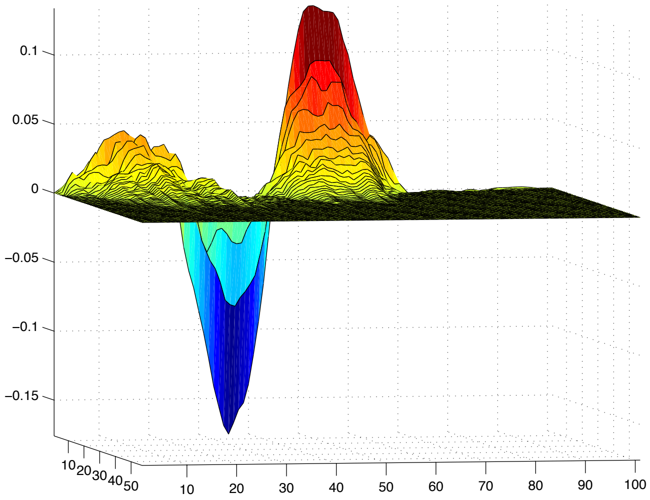
Principal Component Analysis



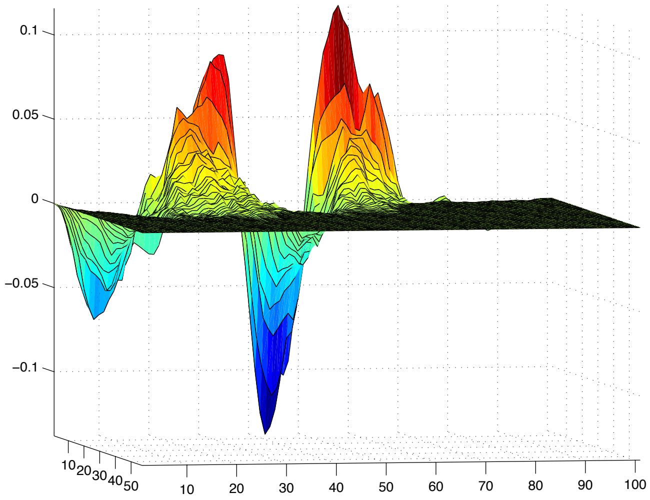
Principal Component Analysis



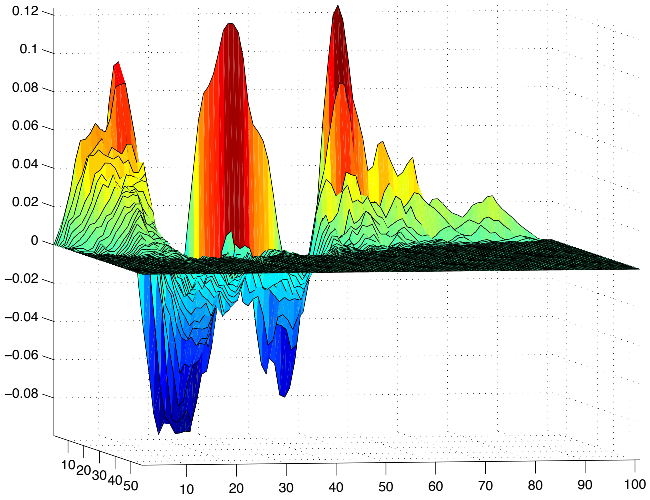
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



Correlation with age

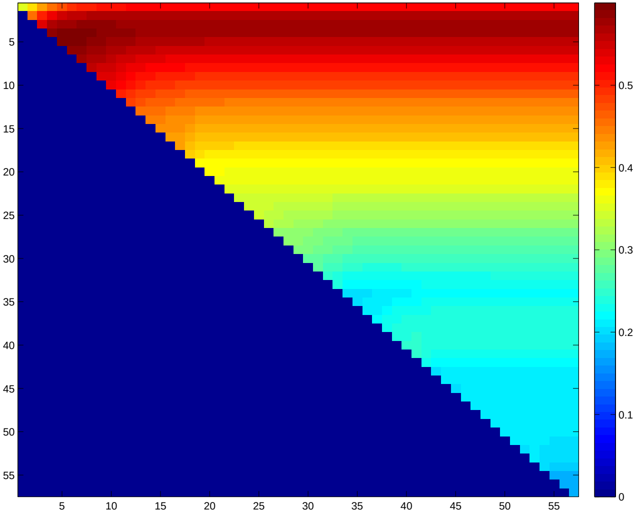
Pearson's correlation coefficient of
age with statistics derived from the brain arteries

Previous study without topology:
Dan Shen et al (2014) $r = 0.25$

Values of r using statistics derived from persistence landscape:

landscapes used	1-norm	first princ comp
$\lambda_1, \ldots, \lambda_{57}$	0.5077	0.5216
$\lambda_2, \ldots, \lambda_{57}$	0.5214	0.5666
$\lambda_5, \ldots, \lambda_5$	0.5582	0.6000

Correlation of age with PCA1 on weighted norms



Summary

- Topology promising tool for analyzing data
- Persistence landscapes easy to combine with standard statistical techniques
- Looking for collaborators

Summary

- Topology promising tool for analyzing data
- Persistence landscapes easy to combine with standard statistical techniques
- Looking for collaborators

Thank you!