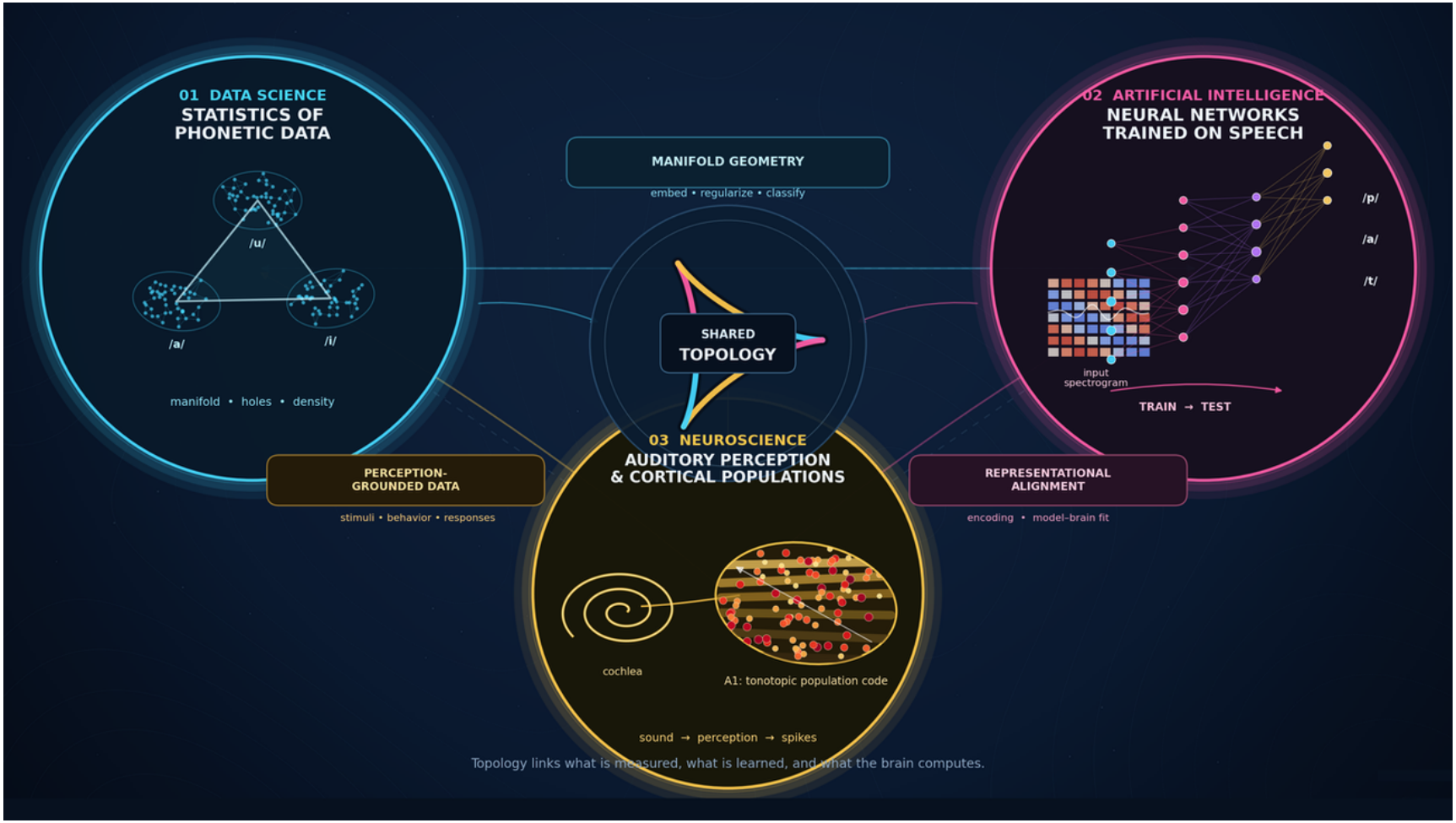
Towards a topological trinity for speech intelligence

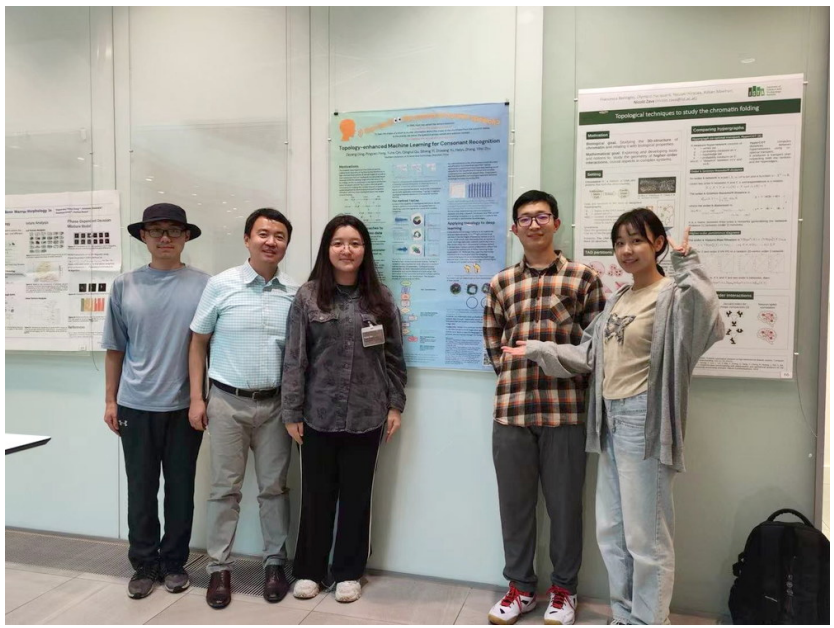
Phonetic statistics • speech-trained neural networks • auditory cortical populations



Yifei Zhu | SUSTech | 2026.11

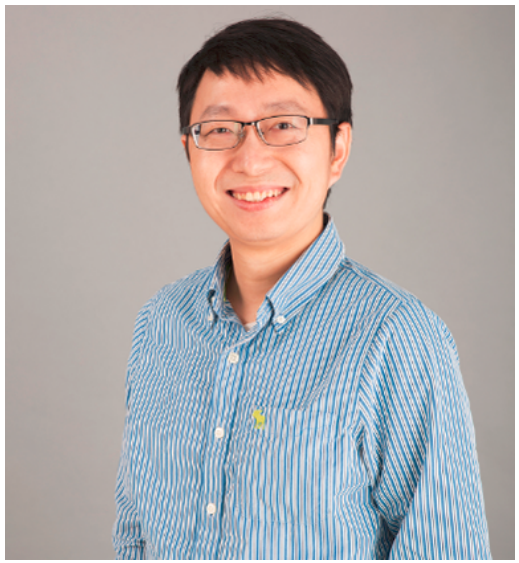
Joint work with –





Zeyang Ding, Pingyao Feng, Qingrui Qu, Siheng Yi,
Zhiwang Yu & Haiyu Zhang, with special thanks to Yuhe Qin

Interdisciplinary research in consultation with –



Dr. Meng Yu

Tencent AI Lab

Speech and audio signal processing



Dr. Yujiao Jennifer Sun

Visual Plasticity Lab
University College London

Visual and auditory neuroscience

Outline and main references

- **Topological Data Analysis & Deep Learning** in a nutshell through applications: the topological trinity of image data
  1. **[Neuroscience]** Detection of topological distributions:
     Singh et al., *Topological analysis of population activity in visual cortex*, 2008.
  2. **[Data Science]** Detection of topological distributions:
     Carlsson, Ishkhanov, de Silva & Zomorodian, *On the local behavior of spaces of natural images*, 2008.
  3. **[Artificial Intelligence]**
     - Detection of topological distributions:
       Brüel Gabrielsson & Carlsson, *Exposition and interpretation of the topology of neural networks*, 2019.
     - Enhancement by topological distributions and dynamics:
       Love, Filippenko, Maroulas & Carlsson, *Topological convolutional layers for deep learning*, 2023.

- **Our research group's work** on the topology of speech data
  - The pipeline of topological time series analysis: Time-Delay Embedding + Persistent Homology
  - **[Data Science & Artificial Intelligence]** Topological features (not yet distributions or dynamics) & enhanced machine learning (topological feature and architecture):
    Feng et al., *Topology-enhanced machine learning for speech signal processing*, 2026.
    Yu et al., *Topology-informed convolution kernels for speech recognition tasks*, preprint.
  - Future directions

Motivation: topology of population activity in visual cortex

In collaboration with mathematicians working in the then-emerging field of topological data analysis (TDA), Ringach and colleagues gave a topological characterization of population activity in primary visual cortex (V1).

Singh et al., Topological analysis of population activity in visual cortex, ***Journal of Vision****, 2008.*

Motivation: topology of population activity in visual cortex

In collaboration with mathematicians working in the then-emerging field of topological data analysis (TDA), Ringach and colleagues gave a topological characterization of population activity in primary visual cortex (V1).

Singh et al., Topological analysis of population activity in visual cortex, ***Journal of Vision****, 2008.*

Basic question. How sensory input and ongoing cortical activity combine to generate a response to a given stimulus?

Motivation: topology of population activity in visual cortex

In collaboration with mathematicians working in the then-emerging field of topological data analysis (TDA), Ringach and colleagues gave a topological characterization of population activity in primary visual cortex (V1).

Singh et al., Topological analysis of population activity in visual cortex, ***Journal of Vision****, 2008.*

Basic question. How sensory input and ongoing cortical activity combine to generate a response to a given stimulus?

Key hypothesis. Natural signals shape the architecture and dynamics of V1.

Motivation: topology of population activity in visual cortex

In collaboration with mathematicians working in the then-emerging field of topological data analysis (TDA), Ringach and colleagues gave a topological characterization of population activity in primary visual cortex (V1).

Singh et al., Topological analysis of population activity in visual cortex, ***Journal of Vision****, 2008.*

Basic question. How sensory input and ongoing cortical activity combine to generate a response to a given stimulus?

Key hypothesis. Natural signals shape the architecture and dynamics of V1.

Computational model. A relatively weak input effectively serves to push the patterns of activity around in a low-dimensional manifold.

Motivation: topology of population activity in visual cortex

In collaboration with mathematicians working in the then-emerging field of topological data analysis (TDA), Ringach and colleagues gave a topological characterization of population activity in primary visual cortex (V1).

Singh et al., Topological analysis of population activity in visual cortex, ***Journal of Vision****, 2008.*

Basic question. How sensory input and ongoing cortical activity combine to generate a response to a given stimulus?

Key hypothesis. Natural signals shape the architecture and dynamics of V1.

Computational model. A relatively weak input effectively serves to push the patterns of activity around in a low-dimensional manifold. It is possible that the structure of this manifold is linked to the structure of the signals one naturally encounters in the environment.

Carlsson, Ishkhanov, de Silva & Zomorodian, On the local behavior of spaces of natural images, ***International Journal of Computer Vision****, 2008.*

Motivation: topology of population activity in visual cortex

In collaboration with mathematicians working in the then-emerging field of topological data analysis (TDA), Ringach and colleagues gave a topological characterization of population activity in primary visual cortex (V1).

Singh et al., Topological analysis of population activity in visual cortex, ***Journal of Vision****, 2008.*

Basic question. How sensory input and ongoing cortical activity combine to generate a response to a given stimulus?

Key hypothesis. Natural signals shape the architecture and dynamics of V1.

Computational model. A relatively weak input effectively serves to push the patterns of activity around in a low-dimensional manifold. It is possible that the structure of this manifold is linked to the structure of the signals one naturally encounters in the environment.

Carlsson, Ishkhanov, de Silva & Zomorodian, On the local behavior of spaces of natural images, ***International Journal of Computer Vision****, 2008.*

We will see analogues of such links in machine learning of visual signals.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

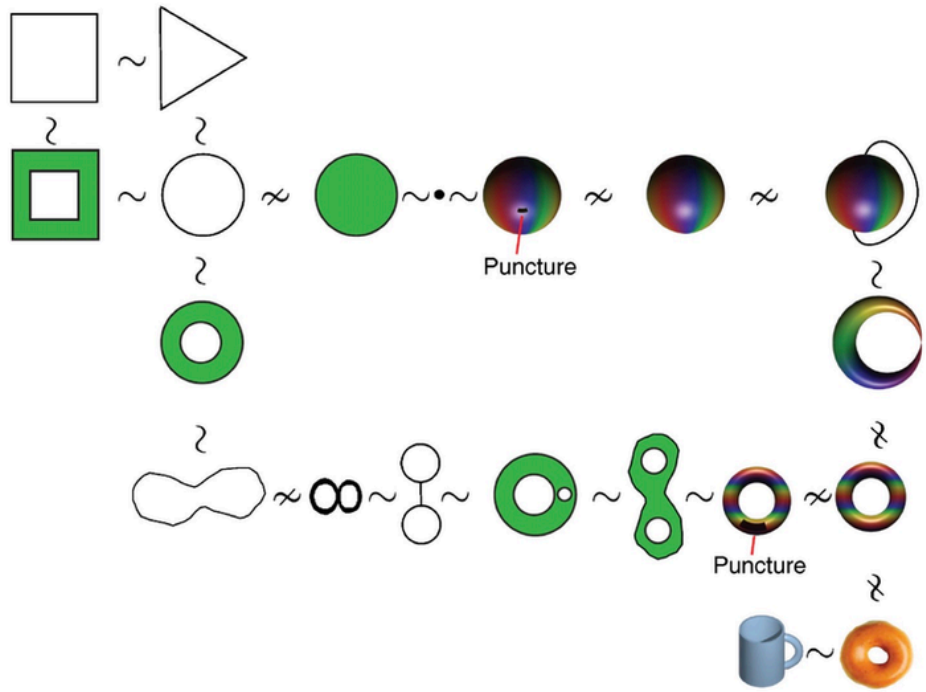


Figure 1. Topological equivalence in rubber-world. The figure illustrates the notion of equivalence by showing several objects (topological spaces) connected by the symbols ∼ when they are equivalent or by ≁ when they are not. The reader should think that all the objects shown are made of an elastic material and visualize the equivalence of two spaces by imagining a deformation between to objects.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)



Source: National Geographic.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)



- *Spontaneous condition*

Source: National Geographic.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)



- *Spontaneous condition*
- *Evoked condition*

Source: National Geographic.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

- *Collect spontaneous and evoked activity segments in lengths of 10 s each*
- *Spike-sort signals from each electrode*

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

- *Collect spontaneous and evoked activity segments in lengths of 10 s each*
- *Spike-sort signals from each electrode*
- *Sub-select a group of 5 neurons that showed the highest firing rates*
- *Bin spikes in 50-ms windows*

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

- *Collect spontaneous and evoked activity segments in lengths of 10 s each*
- *Spike-sort signals from each electrode*
- *Sub-select a group of 5 neurons that showed the highest firing rates*
- *Bin spikes in 50-ms windows*
- *Obtain from each segment 200 points in* $\mathbb{R}^5$

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

- *Collect spontaneous and evoked activity segments in lengths of 10 s each*
- *Spike-sort signals from each electrode*
- *Sub-select a group of 5 neurons that showed the highest firing rates*
- *Bin spikes in 50-ms windows*
- *Obtain from each segment 200 points in* $\mathbb{R}^5$

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

- *Collect spontaneous and evoked activity segments in lengths of 10 s each*
- *Spike-sort signals from each electrode*
- *Sub-select a group of 5 neurons that showed the highest firing rates*
- *Bin spikes in 50-ms windows*
- *Obtain from each segment 200 points in* $\mathbb{R}^5$

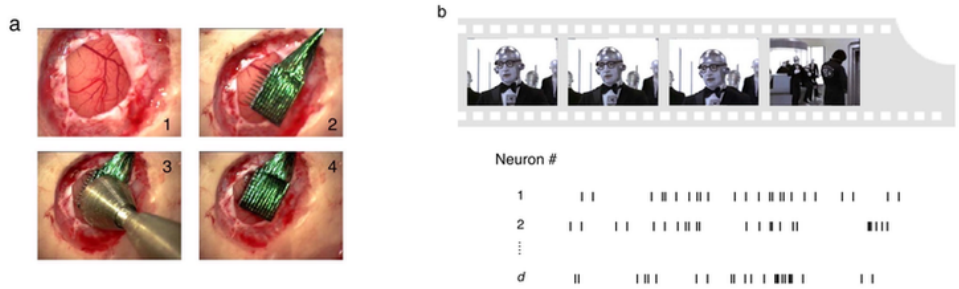


Figure 8. Experimental recordings in primary visual cortex. (a) Insertion sequence of a multi-electrode array into primary visual cortex (Nauhaus & Ringach, 2007). (b) Natural image sequences, sampled from commercial movies, were used to stimulate all receptive fields of neurons isolated by the array. In the spontaneous condition, activity was recorded while both eyes were occluded.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

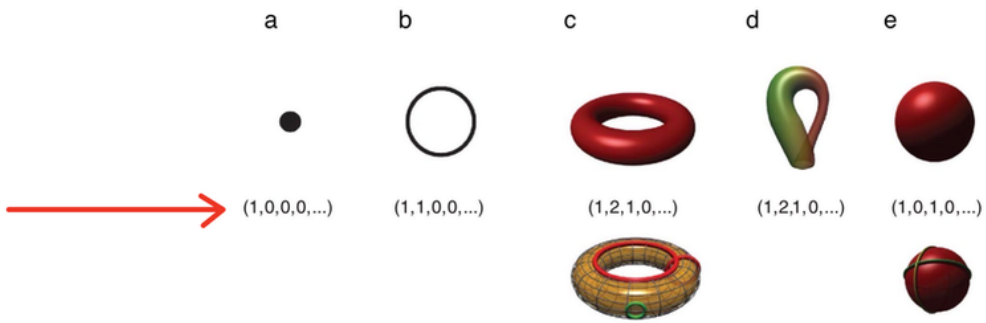


Figure 2. Betti numbers provide a signature of the underlying topology. Illustrated in the figure are five simple objects (topological spaces) together with their Betti number signatures: (a) a point, (b) a circle, (c) a hollow torus, (d) a Klein bottle, and (e) a hollow sphere. For the case of the torus (c), the figure shows three loops on its surface. The red loops are "essential" in that they cannot be shrunk to a point, nor can they be deformed one into the other without tearing the loop. The green loop, on the other hand, can be deformed to a point without any obstruction. For the torus, therefore, we have $b_1 = 2$. For the case of the sphere, the loops shown (and actually all loops on the sphere) can be contracted to points, which is reflected by the fact that $b_1 = 0$. Both the sphere and the torus have $b_2 = 1$, this is due to the fact both surfaces enclose a part of space (a void).

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

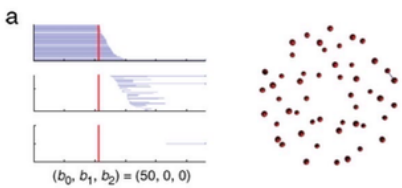


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

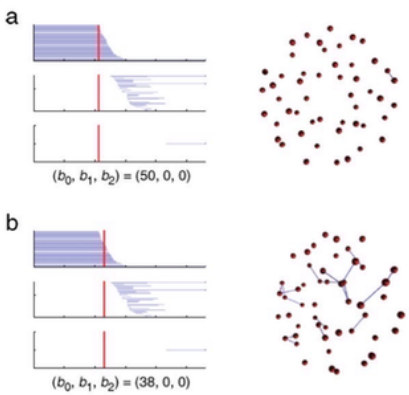


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

ϵ = *measure of closeness (for "clustering")*

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

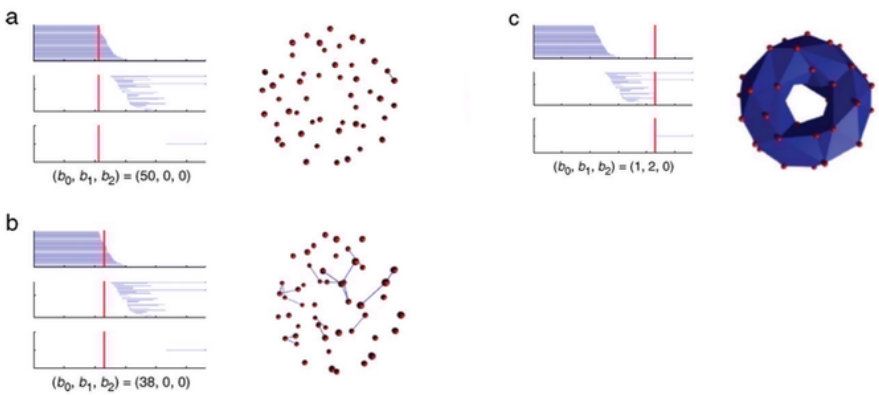


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

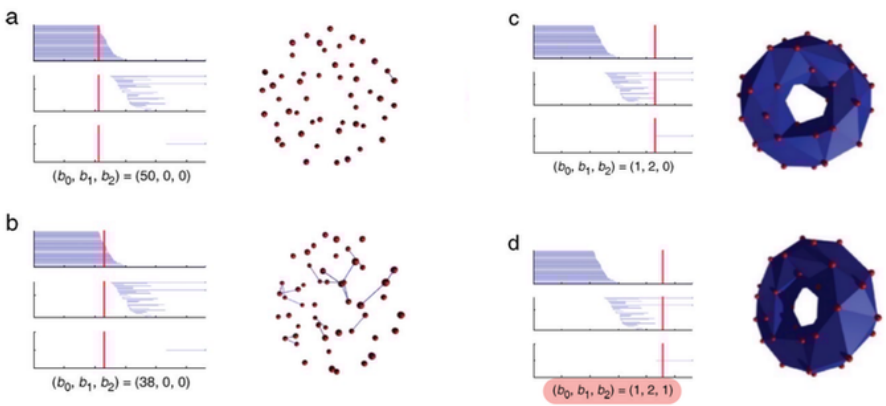


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

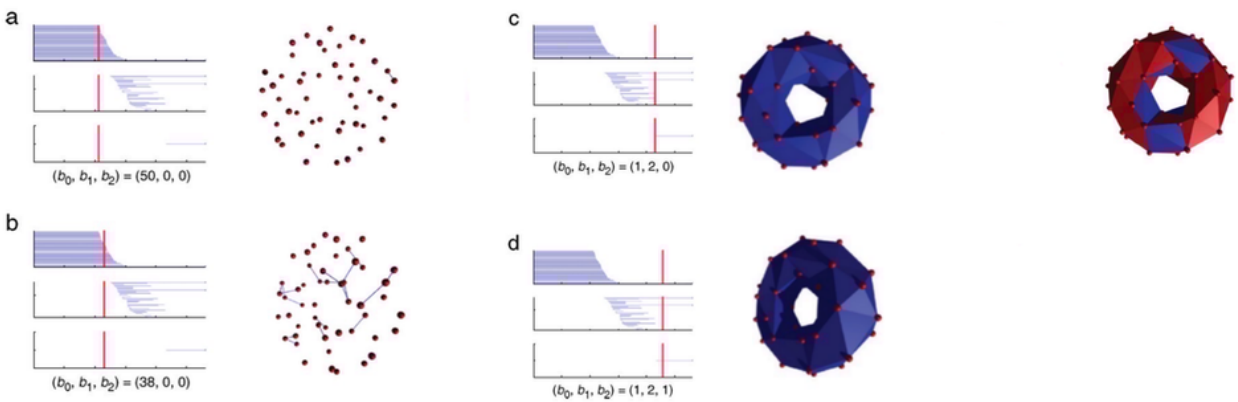


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

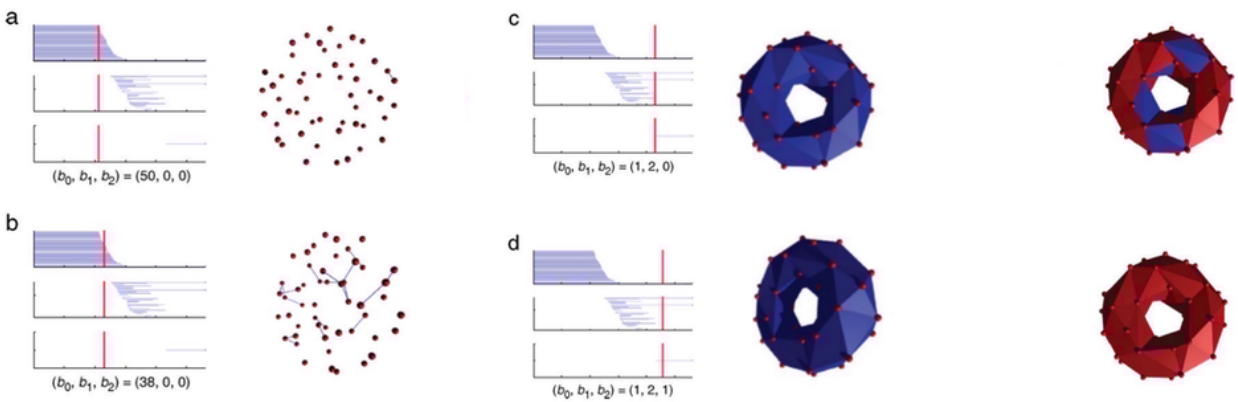


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

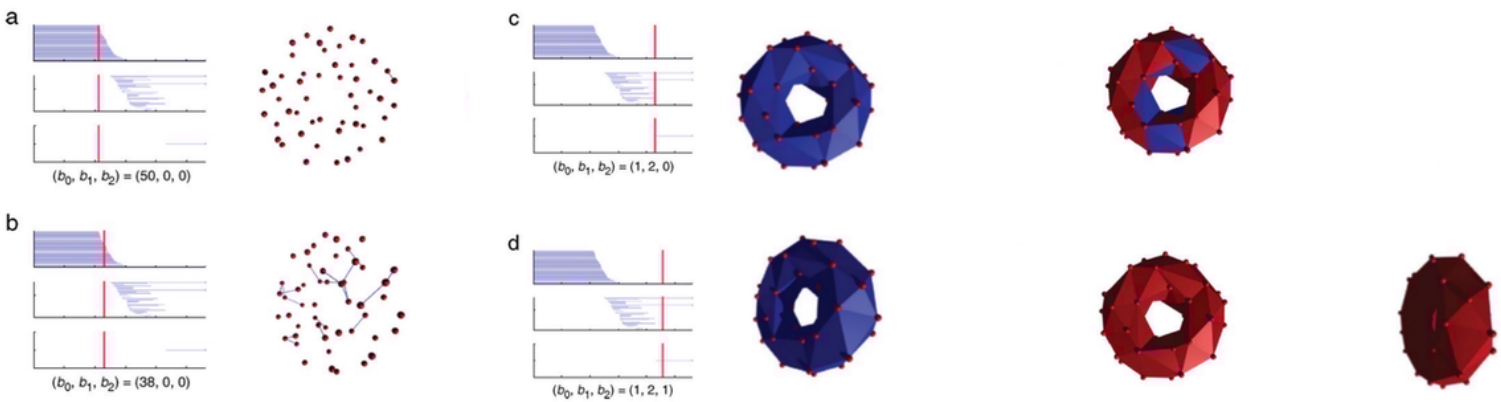


Figure 3. Barcodes and Rips complexes. The figure illustrates the construction of the Rips complex and the generation of barcodes (only the first three Betti numbers are displayed) for 50 points randomly sampled from the surface of a torus. Panels a to d show the barcode "sliced" at different values of ε (the horizontal axis) with the corresponding Rips complexes shown to the right. The corresponding Betti numbers for each level of ε can be obtained by counting the number of horizontal lines crossed by the vertical red line in each graph.

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

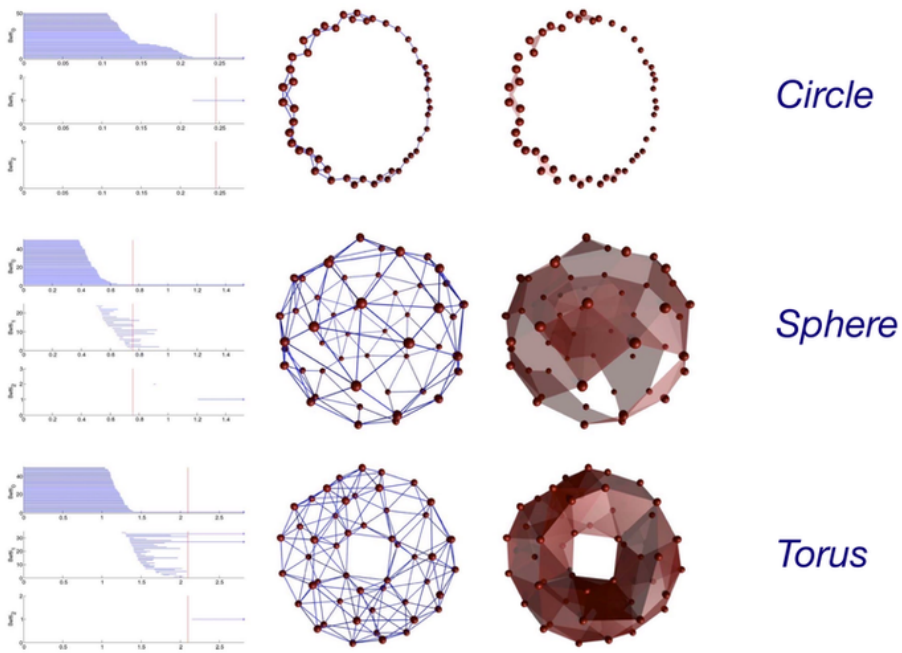
Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)



From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

↓

Results!

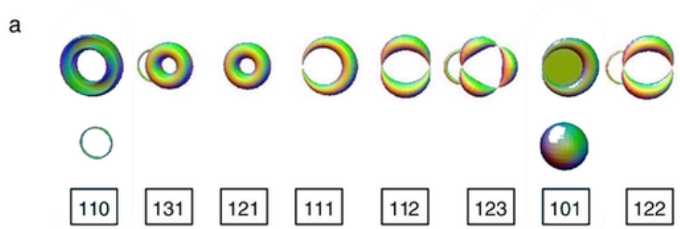


Figure 9. Estimation of topological structure in driven and spontaneous conditions. (a) Ordering of topological signatures observed in our experiments. Each triplet (b_0, b_1, b_2) is shown along an illustration of objects consistent with these signature. (b) Distribution of topological signatures in the spontaneous and natural image stimulation conditions pooled across the three experiments performed. Each row correspond to signatures with a minimum interval length (denoted as the threshold) expressed as a fraction of the covering radius of the data cloud (see Appendix A for the definition of the covering radius).

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

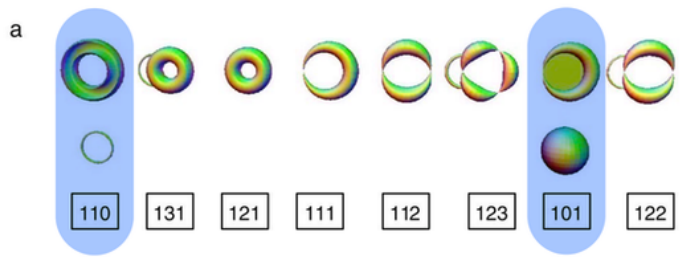
Construction of point clouds

↓

Topological feature extraction (persistent homology)

↓

Results!



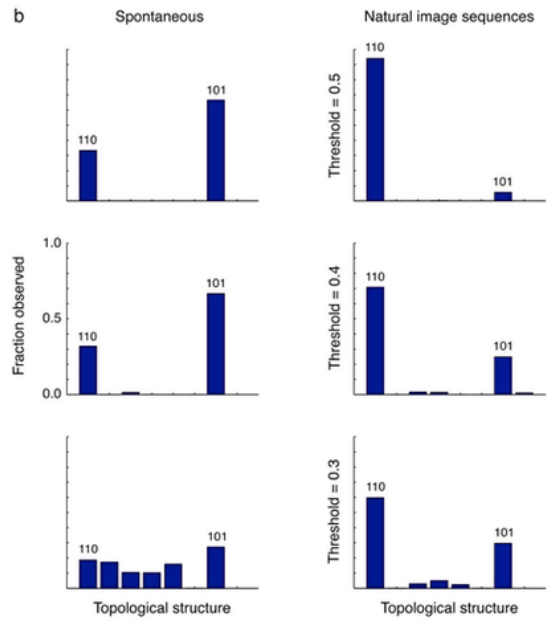


Figure 9. Estimation of topological structure in driven and spontaneous conditions. (a) Ordering of topological signatures observed in our experiments. Each triplet (b_0, b_1, b_2) is shown along an illustration of objects consistent with these signature. (b) Distribution of topological signatures in the spontaneous and natural image stimulation conditions pooled across the three experiments performed. Each row correspond to signatures with a minimum interval length (denoted as the threshold) expressed as a fraction of the covering radius of the data cloud (see Appendix A for the definition of the covering radius).

From Singh et al., 2008.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

↓

Results!

Also validation experiments with simulated signals, robustness, model explanation, etc.

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

↓

Results!

Also validation experiments with simulated signals, robustness, model explanation, etc.

Questions:

- Why does a circle emerge as the topological structure of population activity in V1? Why a sphere?

Motivation: topology of population activity in visual cortex

How to experimentally and computationally detect topological structure in the architecture and dynamics of V1?

Visual stimulation (electrophysiology)

↓

Construction of point clouds

↓

Topological feature extraction (persistent homology)

↓

Results!

Also validation experiments with simulated signals, robustness, model explanation, etc.

Questions:

- Why does a circle emerge as the topological structure of population activity in V1? Why a sphere?
- What about auditory signals? Architecture and dynamics of A1? Analogy or contrast?

Tracing back from perception to signals: natural image statistics

- Lee, Pedersen & Mumford, *The nonlinear statistics of high-contrast patches in natural images*, **International Journal of Computer Vision**, 2003.

- Carlsson, Ishkhanov, de Silva & Zomorodian, *On the local behavior of spaces of natural images*, **International Journal of Computer Vision**, 2008.

Tracing back from perception to signals: natural image statistics

- Lee, Pedersen & Mumford, *The nonlinear statistics of high-contrast patches in natural images*, **International Journal of Computer Vision**, 2003.
- Carlsson, Ishkhanov, de Silva & Zomorodian, *On the local behavior of spaces of natural images*, **International Journal of Computer Vision**, 2008.

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Tracing back from perception to signals: natural image statistics

- Lee, Pedersen & Mumford, *The nonlinear statistics of high-contrast patches in natural images*, **International Journal of Computer Vision**, 2003.
- Carlsson, Ishkhanov, de Silva & Zomorodian, *On the local behavior of spaces of natural images*, **International Journal of Computer Vision**, 2008.

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Each pixel has a grayscale value, can be thought of as a real number (in reality, takes one of 256 values).

Tracing back from perception to signals: natural image statistics

- Lee, Pedersen & Mumford, *The nonlinear statistics of high-contrast patches in natural images*, **International Journal of Computer Vision**, 2003.
- Carlsson, Ishkhanov, de Silva & Zomorodian, *On the local behavior of spaces of natural images*, **International Journal of Computer Vision**, 2008.

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Each pixel has a grayscale value, can be thought of as a real number (in reality, takes one of 256 values).

Typical camera uses tens of thousands of pixels, so images lie in a very high-dimensional space, call it pixel space, $\mathscr{P}$.

Tracing back from perception to signals: natural image statistics

- Lee, Pedersen & Mumford, *The nonlinear statistics of high-contrast patches in natural images*, **International Journal of Computer Vision**, 2003.
- Carlsson, Ishkhanov, de Silva & Zomorodian, *On the local behavior of spaces of natural images*, **International Journal of Computer Vision**, 2008.

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Each pixel has a grayscale value, can be thought of as a real number (in reality, takes one of 256 values).

Typical camera uses tens of thousands of pixels, so images lie in a very high-dimensional space, call it pixel space, $\mathscr{P}$.

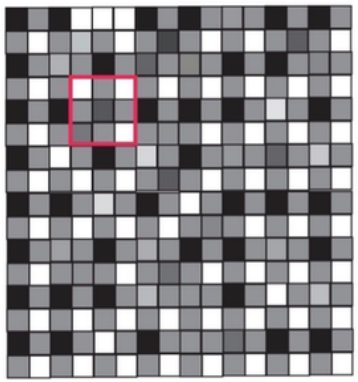
Mumford asks: What can be said about the set of images $\mathscr{I} \subset \mathscr{P}$ one obtains when one takes many images with a digital camera?

Tracing back from perception to signals: natural image statistics

Lee, Mumford, Pedersen: Useful to study *local* structure of images statistically.

Tracing back from perception to signals: natural image statistics

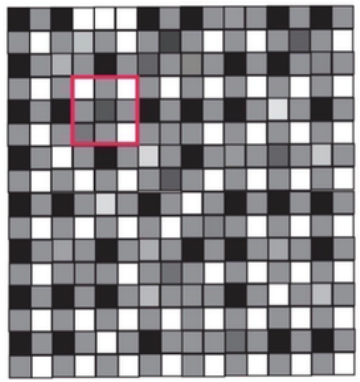
Lee, Mumford, Pedersen: Useful to study *local* structure of images statistically.



3 x 3 patches in images

Tracing back from perception to signals: natural image statistics

Lee, Mumford, Pedersen: Useful to study *local* structure of images statistically.



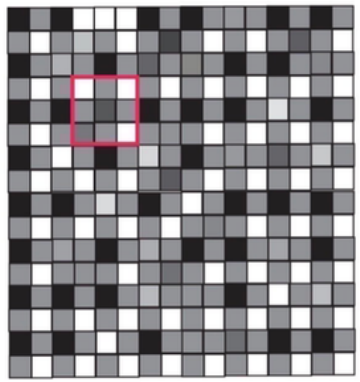
3 x 3 patches in images

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$.

Tracing back from perception to signals: natural image statistics

Lee, Mumford, Pedersen: Useful to study *local* structure of images statistically.



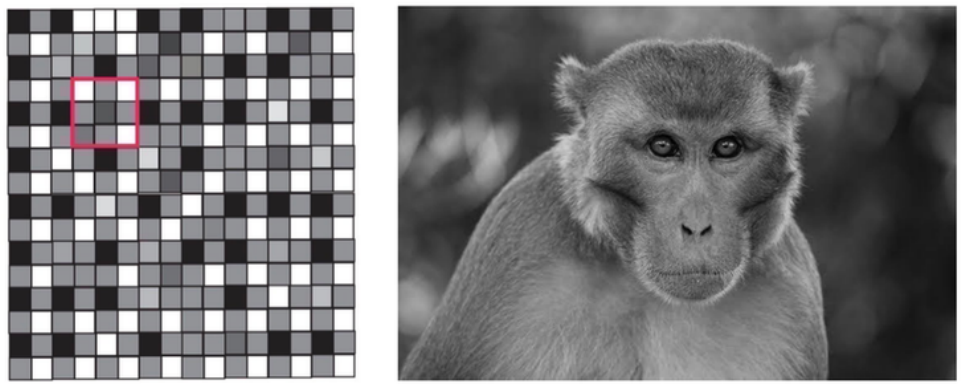
3 x 3 patches in images

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$.
2. Most patches will be nearly constant, or *low* contrast, because of the presence of regions of solid shading in most images.

Tracing back from perception to signals: natural image statistics

Lee, Mumford, Pedersen: Useful to study *local* structure of images statistically.



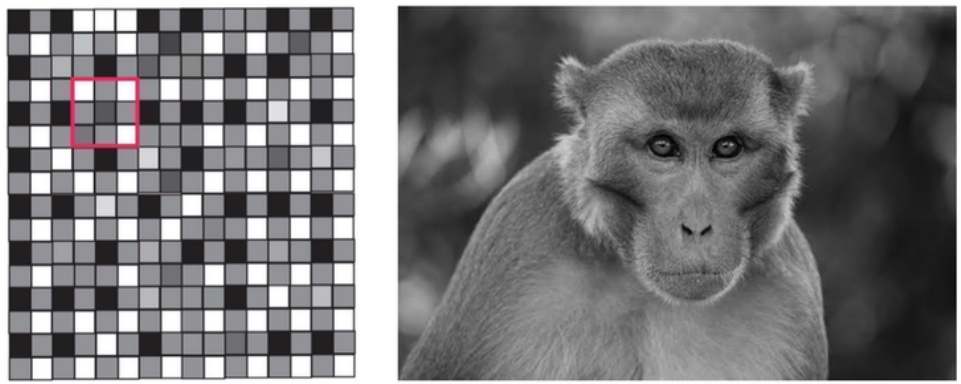
3 x 3 patches in images

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$.
2. Most patches will be nearly constant, or *low* contrast, because of the presence of regions of solid shading in most images.

Tracing back from perception to signals: natural image statistics

Lee, Mumford, Pedersen: Useful to study *local* structure of images statistically.



3 x 3 patches in images

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$.
2. Most patches will be nearly constant, or *low* contrast, because of the presence of regions of solid shading in most images.
3. Low contrast will dominate statistics, not interesting. *High* contrast patches delineate profiles.

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

- Collect approximately 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

- Collect approximately 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.

- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity = 0.

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

- Collect approximately 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity = 0.
- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$.

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

- Collect approximately 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity = 0.
- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$.
- Normalize contrast by dividing by the norm, so obtain patches with norm = 1.

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

- Collect approximately 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity = 0.
- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$.
- Normalize contrast by dividing by the norm, so obtain patches with norm = 1.
- So, data now lie on a 7-dimensional sphere, $\cong S^7$.

Tracing back from perception to signals: natural image statistics

Lee, Mumford & Pedersen study only high contrast patches:

- Collect approximately 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity = 0.
- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$.
- Normalize contrast by dividing by the norm, so obtain patches with norm = 1.
- So, data now lie on a 7-dimensional sphere, $\cong S^7$.

Result: Point cloud data $\mathscr{M}$ lying on a sphere in $\mathbb{R}^8$.

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$.

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze?

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$.

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

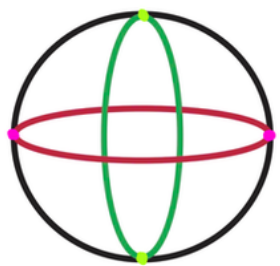
First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

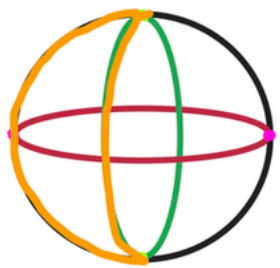
First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

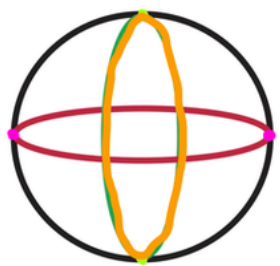
First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

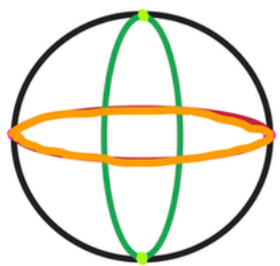
First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

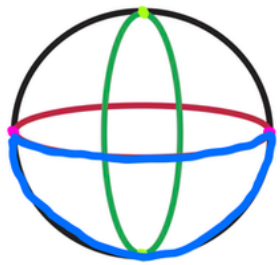
First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

Tracing back from perception to signals: natural image statistics

Carlsson, Ishkhanov, de Silva & Zomorodian analyze it with persistent homology to understand it qualitatively.

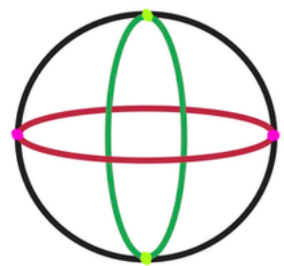
First observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathscr{M}$. However, density of points varies a great deal from region to region.

How to analyze? Set *thresholds* for $\mathscr{M}$. Define $\mathscr{M}[T] \subset \mathscr{M}$ by

$$\mathscr{M}[T] = \{x \mid x \text{ is in } T\text{'th percentile of densest points}\}$$

By computing the persistent homology of these $\mathscr{M}[T]$'s, they reveal

1. 5×10^4 points, $T = 25$:
   There are 5 independent 1-dimensional cycles on $\mathscr{M}[T]$.
   Red and green circles don't touch, each touches black circle.



Three-circle model

2. 4.5×10^6 points, $T = 10$:
   There are one 0D cycle (connected), two 1D cycles (loops),
   and one 2D cycle (surface), i.e., topological signature $= (1, 2, 1)$.
   Is there a surface in which this picture fits?

Tracing back from perception to signals: natural image statistics

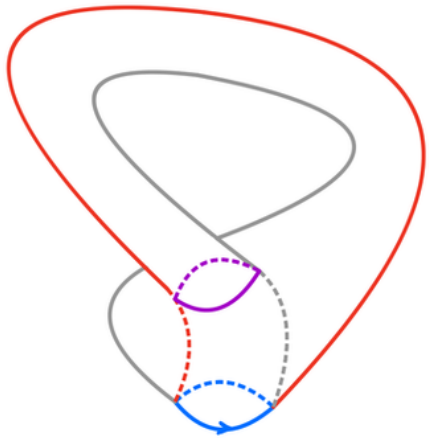
Klein bottle!

Tracing back from perception to signals: natural image statistics

Klein bottle!

Visualization:



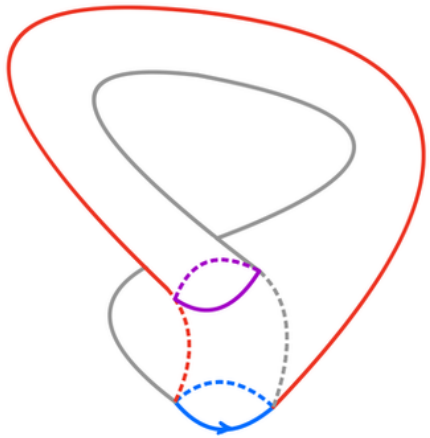


Tracing back from perception to signals: natural image statistics

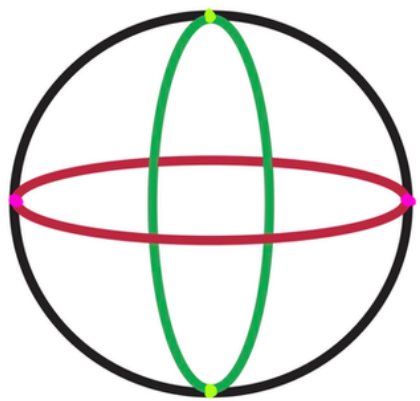
Klein bottle!

Visualization:



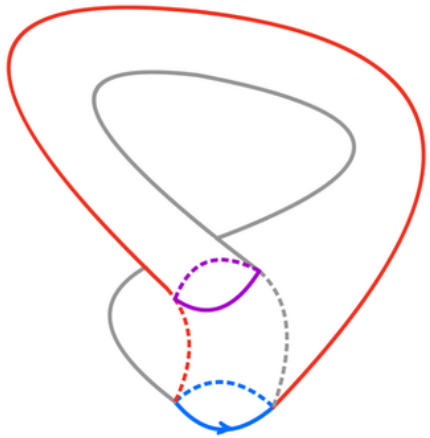


Three circles fit naturally inside the Klein bottle?



Tracing back from perception to signals: natural image statistics

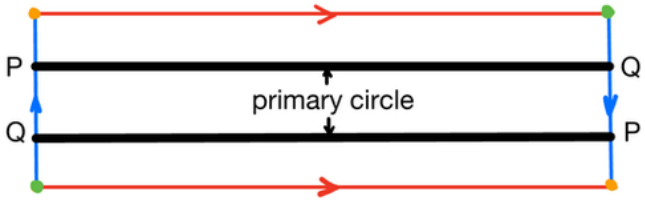
Klein bottle!

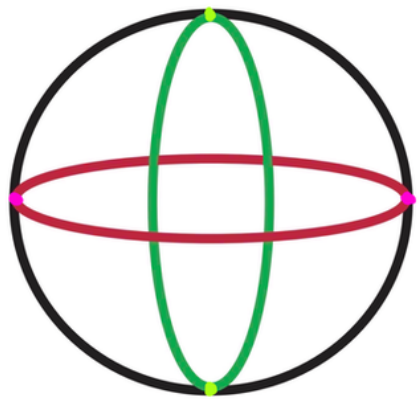


Visualization:



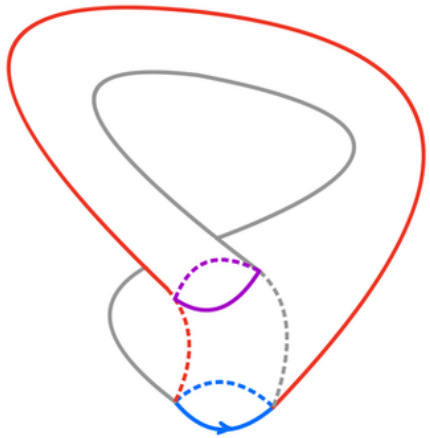
Three circles fit naturally inside the Klein bottle?





Tracing back from perception to signals: natural image statistics

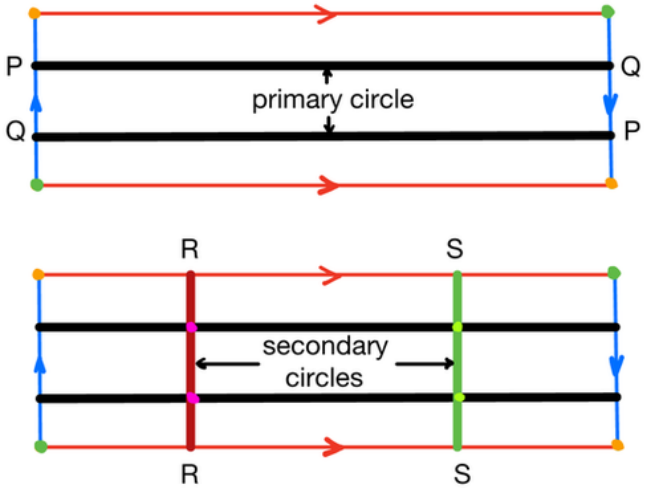
Klein bottle!

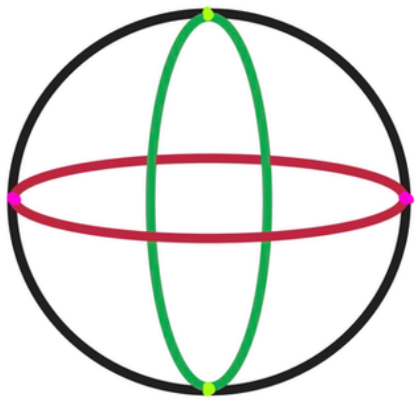


Visualization:



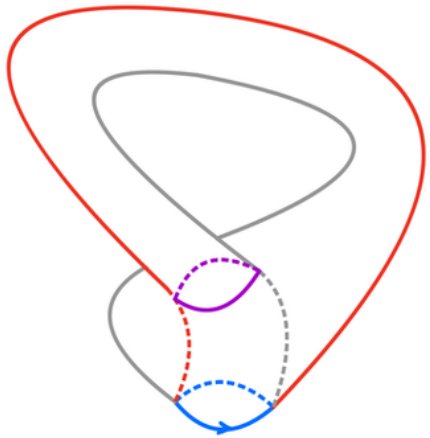
Three circles fit naturally inside the Klein bottle?





Tracing back from perception to signals: natural image statistics

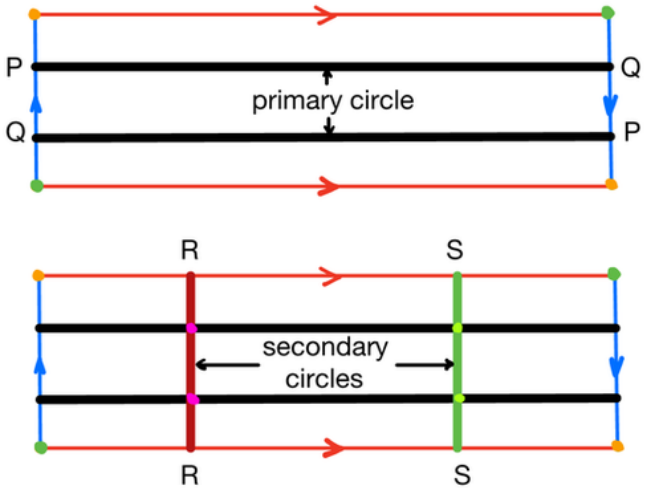
Klein bottle!

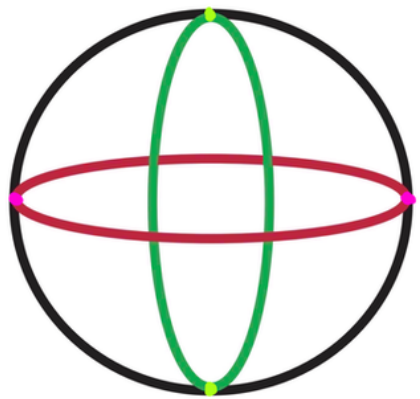


Visualization:



Three circles fit naturally inside the Klein bottle?





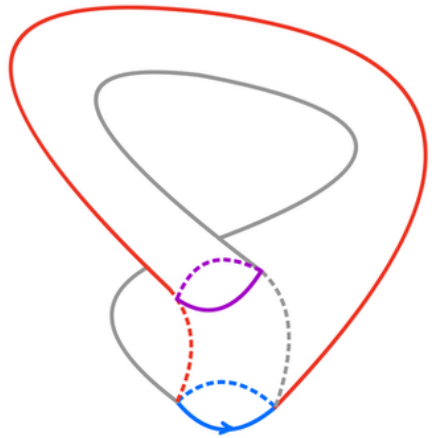
Klein bottle emerges as where local data from digital camera images distribute.

Tracing back from perception to signals: natural image statistics

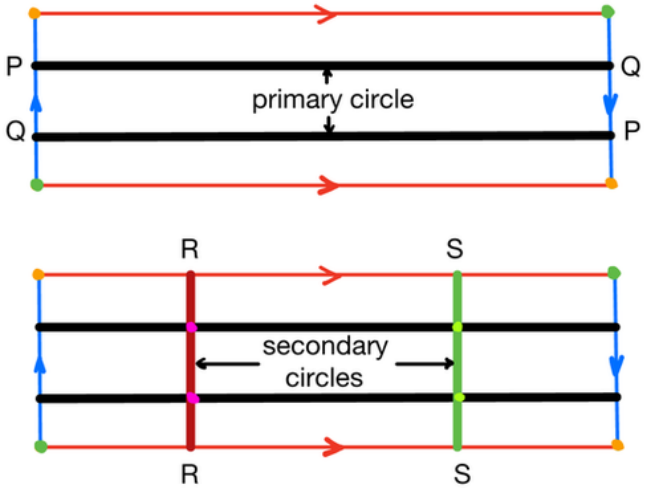
Klein bottle!

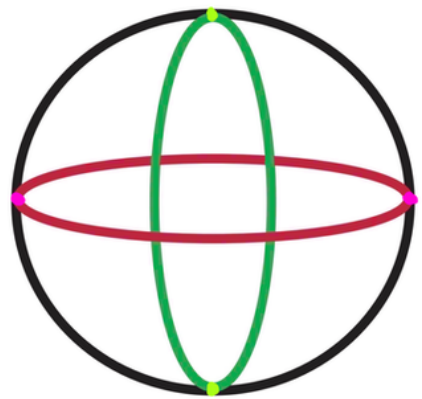
Visualization:

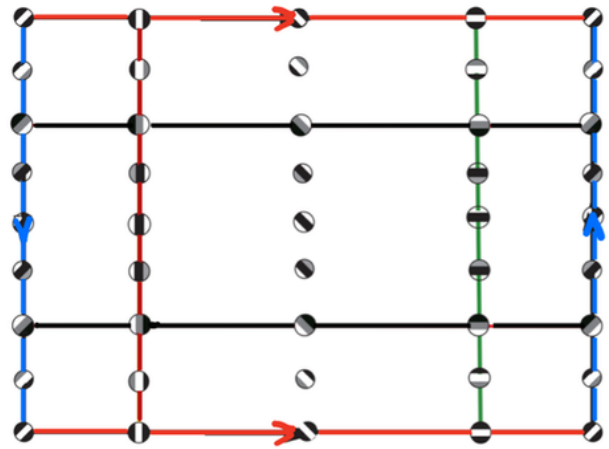




Three circles fit naturally inside the Klein bottle?





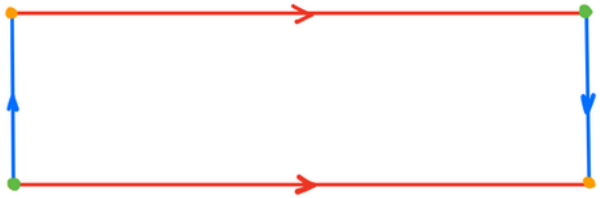


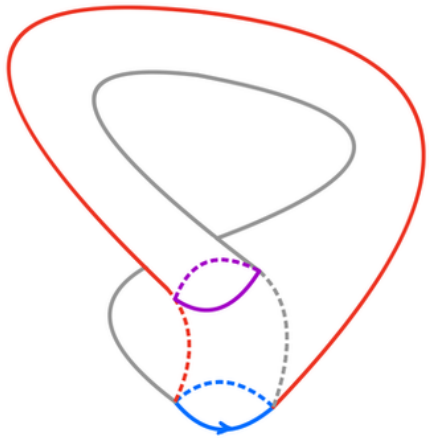
Klein bottle emerges as where local data from digital camera images distribute.

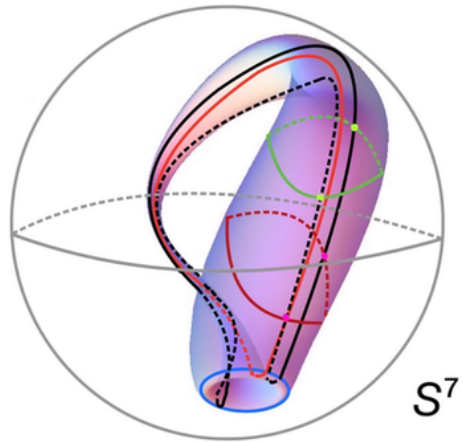
Tracing back from perception to signals: natural image statistics

Klein bottle!

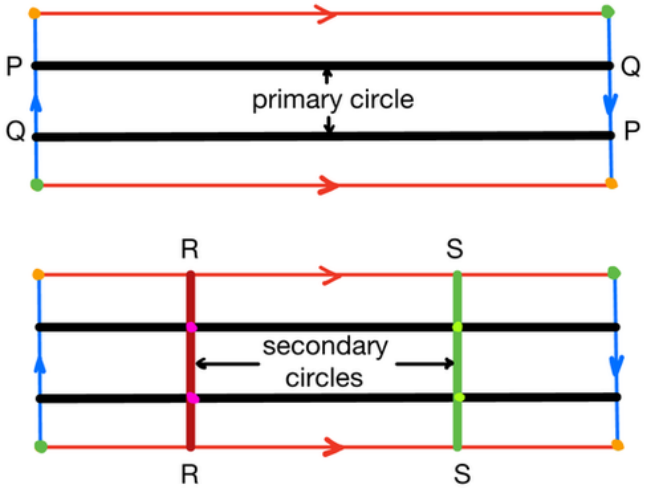
Visualization:

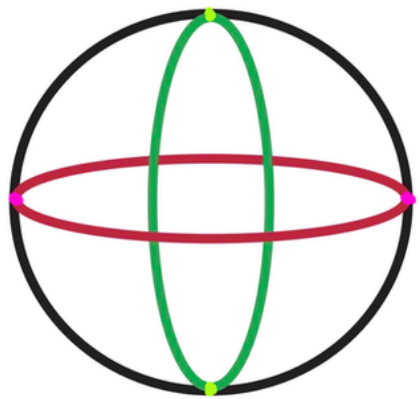


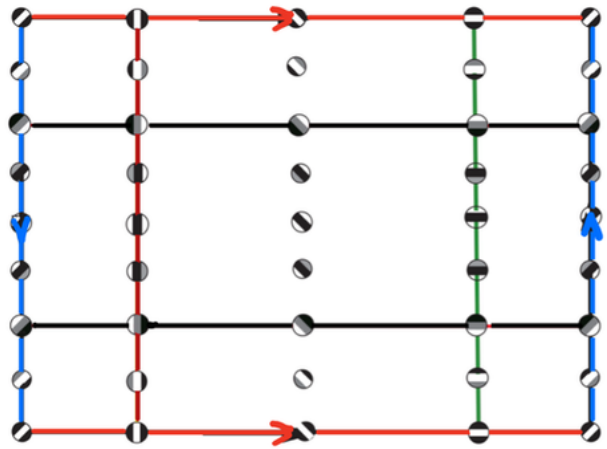




Three circles fit naturally inside the Klein bottle?







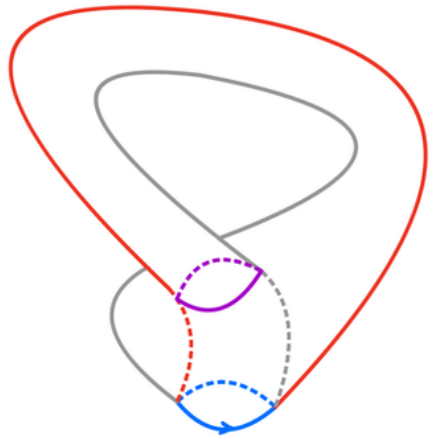
Klein bottle emerges as where local data from digital camera images distribute.

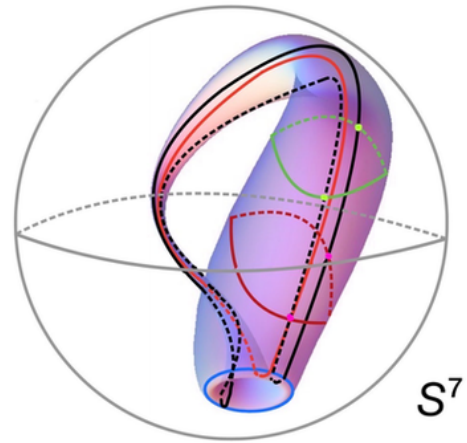
Tracing back from perception to signals: natural image statistics

Klein bottle!

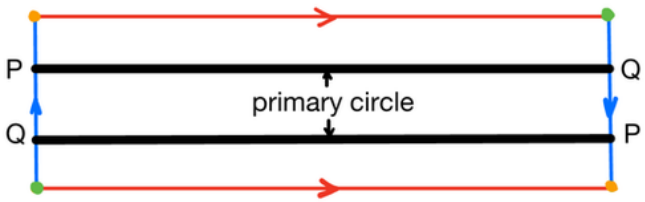
Visualization:

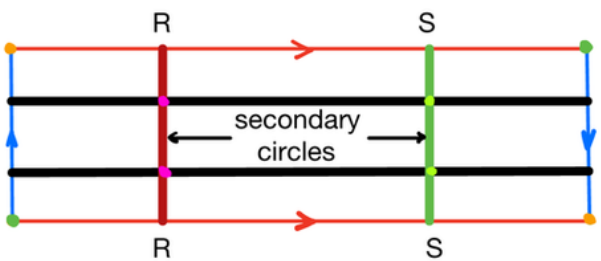


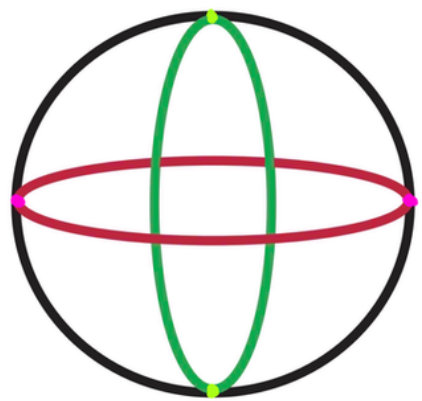


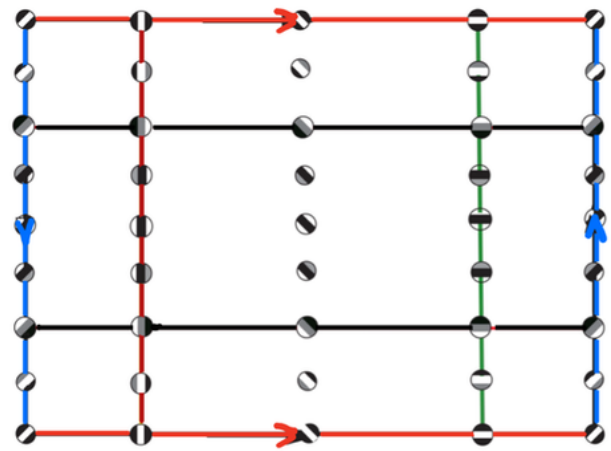


Three circles fit naturally inside the Klein bottle?





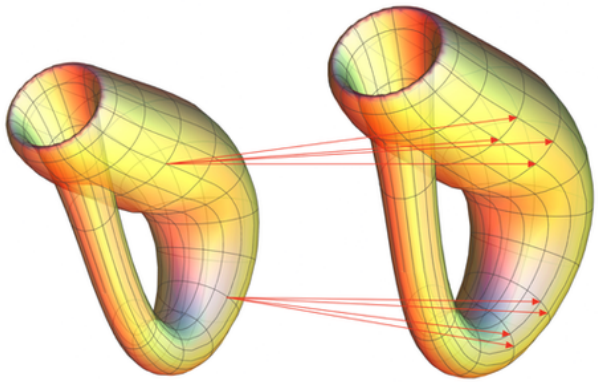




With hindsight, just as model explanation for the neuroscience experiments above, there is indeed a theoretical reason for this distribution over the Klein bottle, rather than over some other 2D manifolds such as a torus.

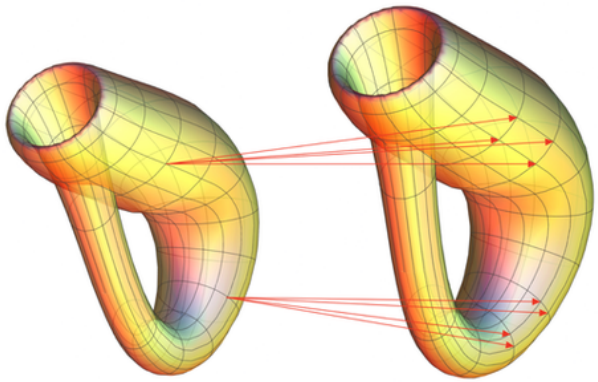
From image data analysis to deep learning

A decade later, Love, Filippenko, Maroulas & Carlsson have made the Klein bottle distribution a topological input for designing convolutional layers in neural networks that learn image data.



From image data analysis to deep learning

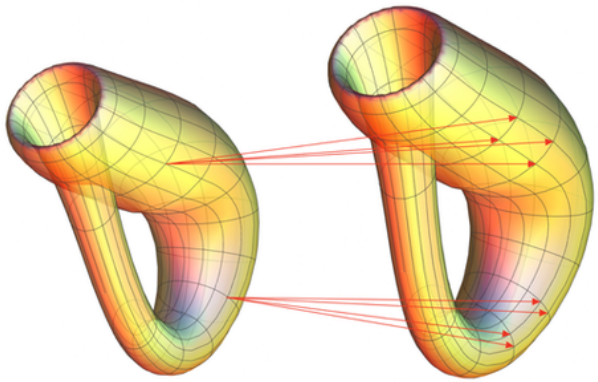
A decade later, Love, Filippenko, Maroulas & Carlsson have made the Klein bottle distribution a topological input for designing convolutional layers in neural networks that learn image data.



Moreover, they have incorporated the tangent bundle of a Klein bottle into TCNNs for learning video data.

From image data analysis to deep learning

A decade later, Love, Filippenko, Maroulas & Carlsson have made the Klein bottle distribution a topological input for designing convolutional layers in neural networks that learn image data.



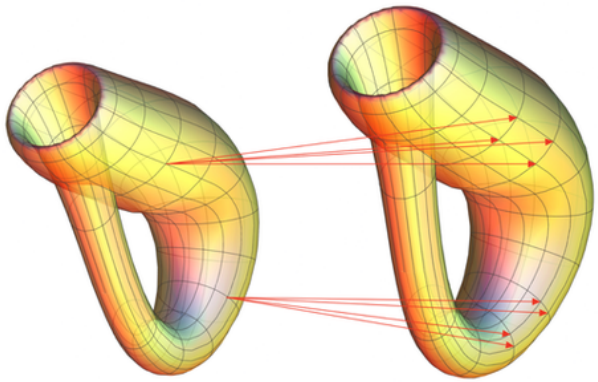
Moreover, they have incorporated the tangent bundle of a Klein bottle into TCNNs for learning video data.

Both learnings achieved higher accuracies with smaller training sets.

- Love, Filippenko, Maroulas & Carlsson, *Topological convolutional layers for deep learning*, **Journal of Machine Learning Research**, 2023.

From image data analysis to deep learning

A decade later, Love, Filippenko, Maroulas & Carlsson have made the Klein bottle distribution a topological input for designing convolutional layers in neural networks that learn image data.



Moreover, they have incorporated the tangent bundle of a Klein bottle into TCNNs for learning video data.

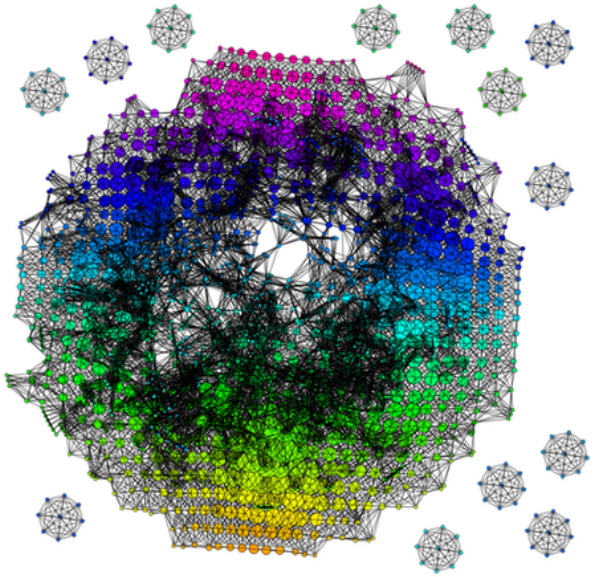
Both learnings achieved higher accuracies with smaller training sets.

- Love, Filippenko, Maroulas & Carlsson, *Topological convolutional layers for deep learning*, **Journal of Machine Learning Research**, 2023.

- Brüel Gabrielsson & Carlsson, *Exposition and interpretation of the topology of neural networks*, **IEEE International Conference on Machine Learning and Applications**, 2019.

From image data analysis to deep learning

Topology of convolutional neural networks: Emergence of cycles during a training process

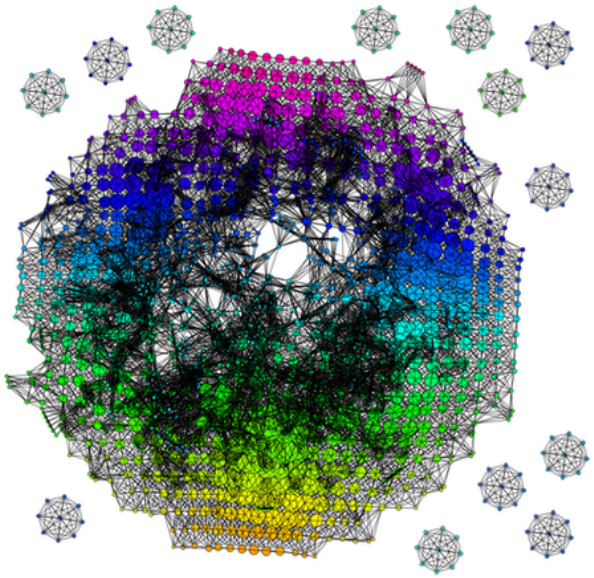


Untrained

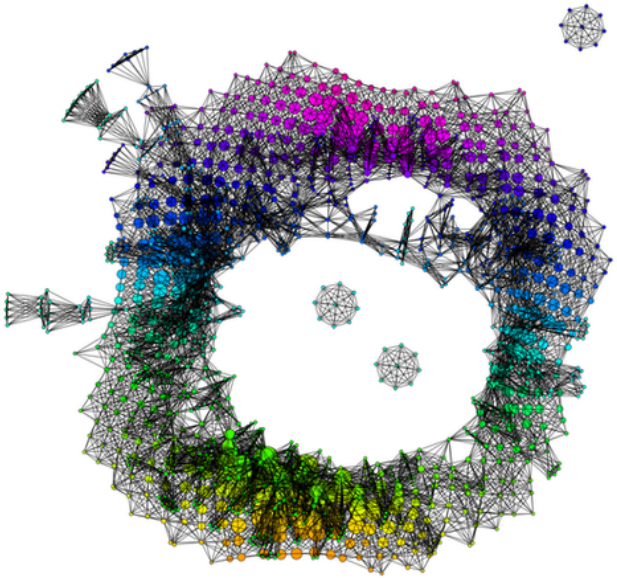
Reproduced using GUDHI,
after Brüel Gabrielsson and Carlsson

From image data analysis to deep learning

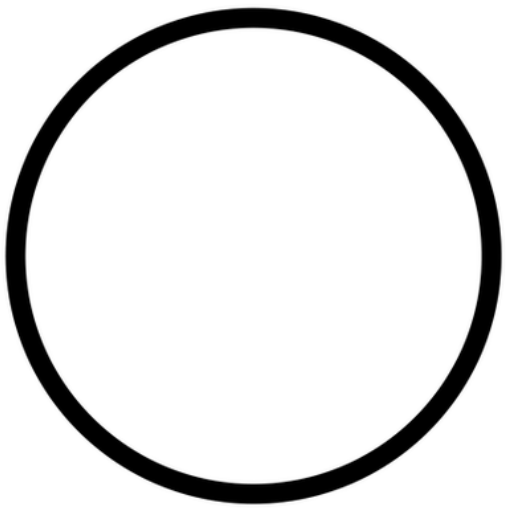
Topology of convolutional neural networks: Emergence of cycles during a training process



Untrained



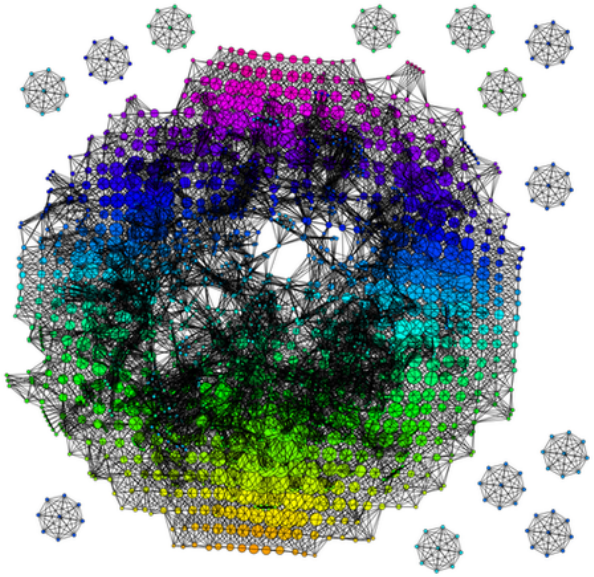
After 5 epochs



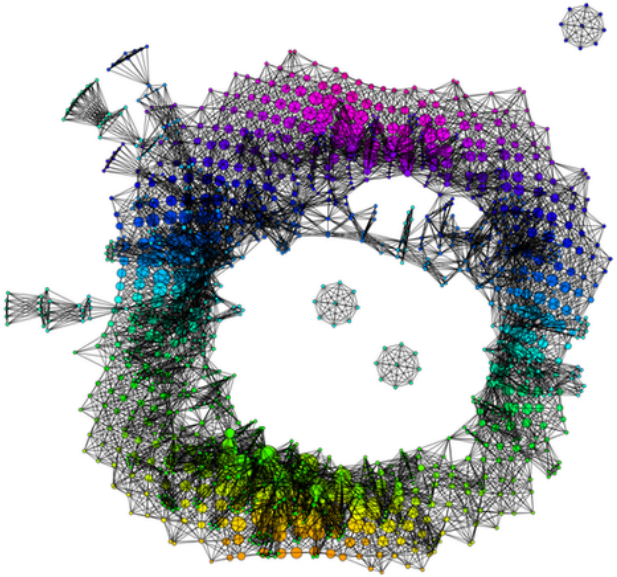
Reproduced using GUDHI,
after Brüel Gabrielsson and Carlsson

From image data analysis to deep learning

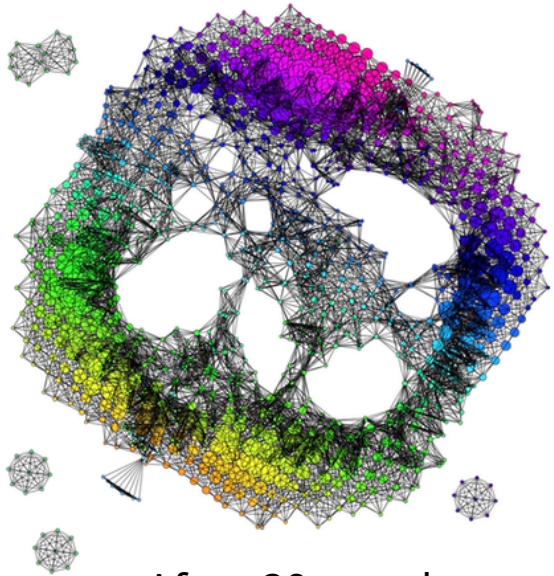
Topology of convolutional neural networks: Emergence of cycles during a training process



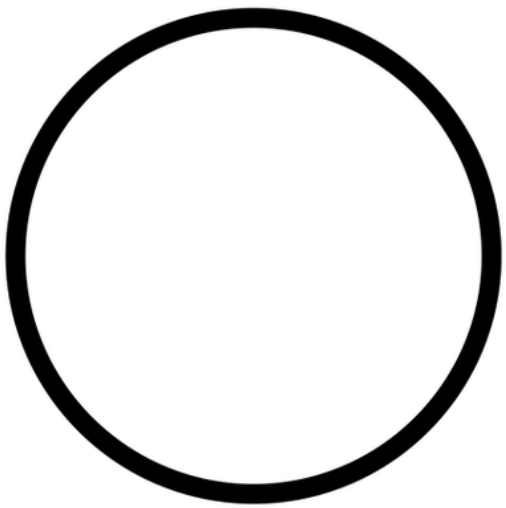
Untrained

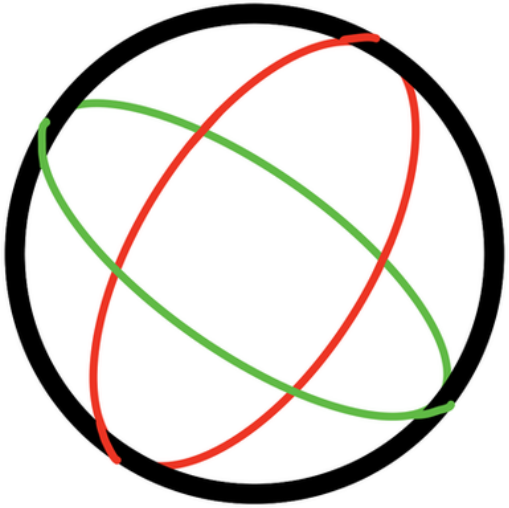


After 5 epochs



After 20 epochs





Reproduced using GUDHI, after Brüel Gabrielsson and Carlsson

From visual to auditory perception

In 1966, Mark Kac asked the famous question:

Can you hear the shape of a drum?

From visual to auditory perception

In 1966, Mark Kac asked the famous question:

Can you hear the shape of a drum?

To hear the shape of a drum is to deduce information about the shape of the drumhead from the sound it makes, using mathematical theory.

From visual to auditory perception

In 1966, Mark Kac asked the famous question:

Can you hear the shape of a drum?

To hear the shape of a drum is to deduce information about the shape of the drumhead from the sound it makes, using mathematical theory.

Now, we mirror the question across senses and address instead:

Can you see the sound of a human speech?

From visual to auditory perception

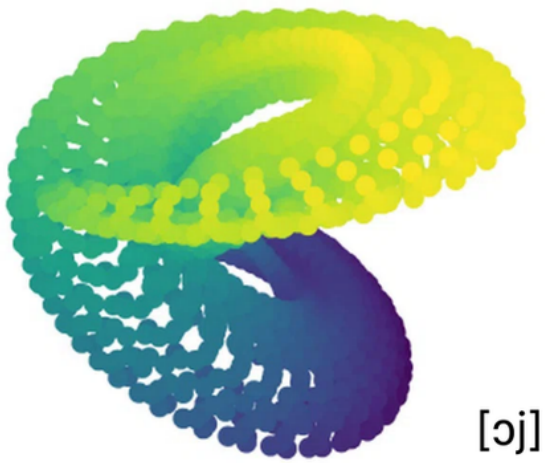
In 1966, Mark Kac asked the famous question:

Can you hear the shape of a drum?

To hear the shape of a drum is to deduce information about the shape of the drumhead from the sound it makes, using mathematical theory.

Now, we mirror the question across senses and address instead:

Can you see the sound of a human speech?



Overview: context & summary

Topological speech (and audio) signal processing

Overview: context & summary

Topological speech (and audio) signal processing

time series data

Overview: context & summary

Topological speech (and audio) signal processing

time series data

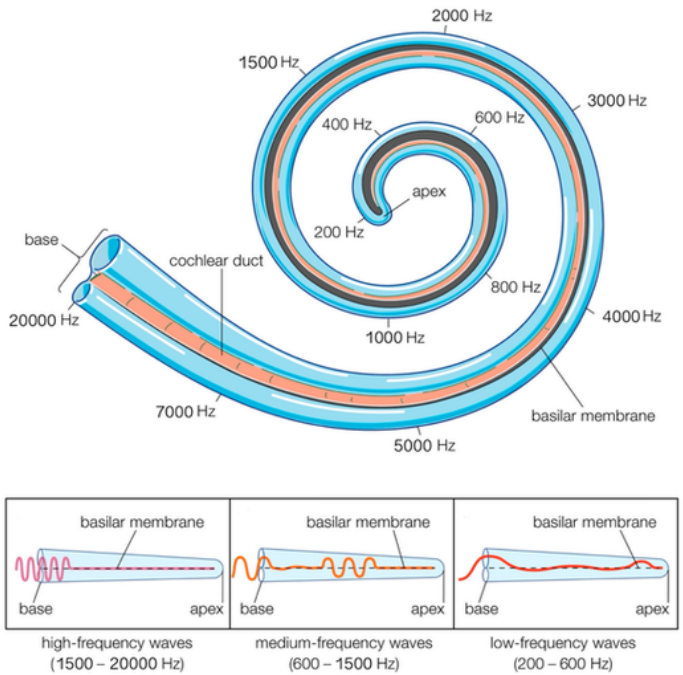
one of the essential components of AI

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering:



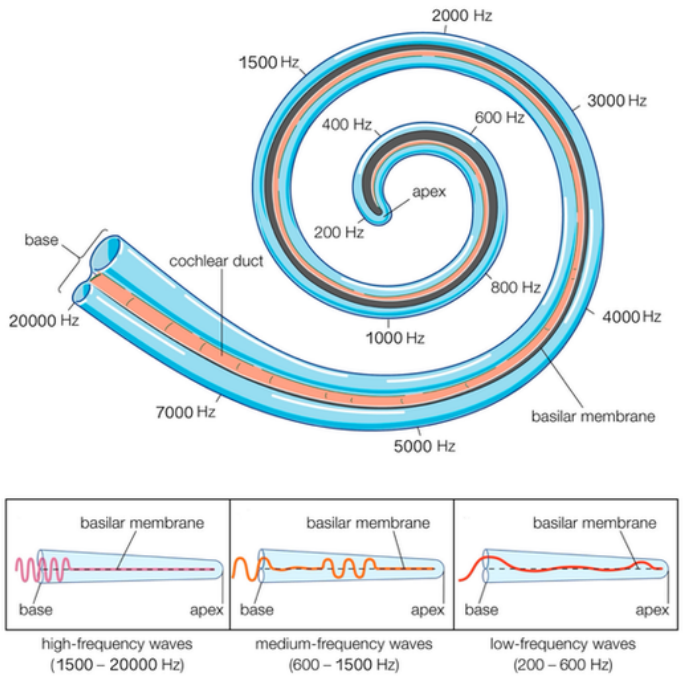
Adapted from Encyclopædia Britannica

Distribution of frequencies along the basilar membrane of the cochlea, which functions as a natural Fourier analysis device

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT

short-time Fourier transform



Adapted from Encyclopædia Britannica

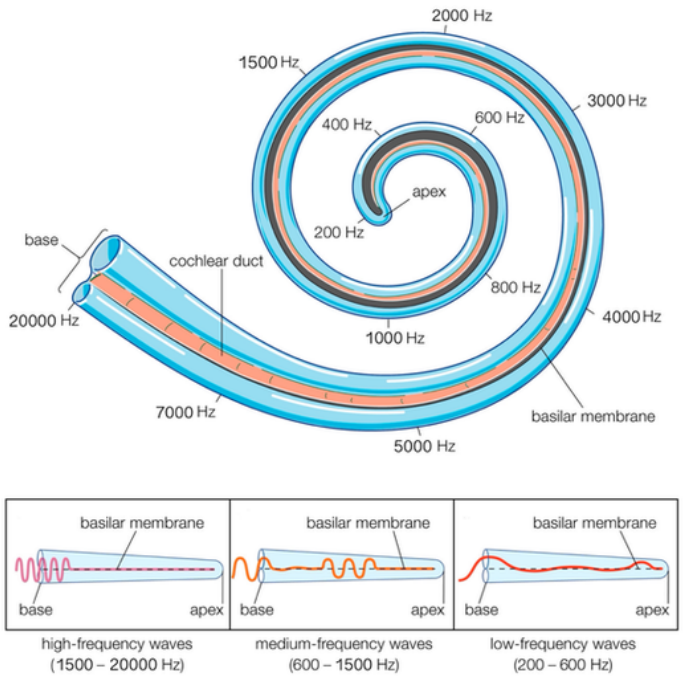
Distribution of frequencies along the basilar membrane of the cochlea, which functions as a natural Fourier analysis device

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

mel-frequency cepstral coefficients

short-time Fourier transform



Adapted from Encyclopædia Britannica

Distribution of frequencies along the basilar membrane of the cochlea, which functions as a natural Fourier analysis device

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML *machine learning*

topological data analysis

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap
2. TopNN
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap *capability of capturing topological structures of data*
2. TopNN
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap *capability of capturing topological structures of data*
2. TopNN *topology-enhanced neural network*
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap — *capability of capturing topological structures of data*
2. TopNN — *topology-enhanced neural network*
3. TopKer — *topology-informed convolution kernel*

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN: outperforms, accuracy + convergence of loss function + steadiness + robustness against noise
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN: outperforms, accuracy + convergence of loss function + steadiness + robustness against noise
3. TopKer

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN: outperforms, accuracy + convergence of loss function + steadiness + robustness against noise
3. TopKer: superior performance

phoneme recognition

Overview: context & summary

Topological speech (and audio) signal processing, beyond direct biomimetic engineering: topological features vs. STFT/MFCC

Combination of TDA to ML:

1. TopCap: stands in comparison, datasets + models
2. TopNN: outperforms, accuracy + convergence of loss function + steadiness + robustness against noise
3. TopKer: superior performance + cross-domain adaptability

phoneme recognition *other audio and visual recognition tasks*

Periodic phenomena: a motivating example

Let $T^2 = (\mathbb{R}/\mathbb{Z})^2$ be the 2D torus. Consider the dynamical system given by

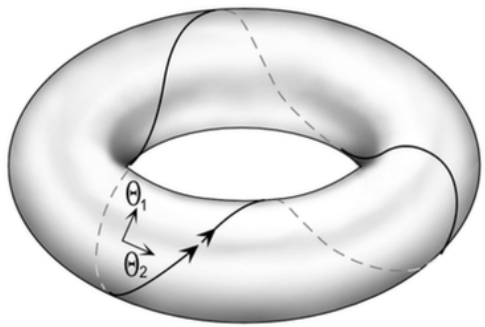
$$\begin{aligned} \Phi_\sigma : T^2 \times \mathbb{R} &\to T^2 \\ \big((a,b),t\big) &\mapsto (a+t,\ b+\sigma t) \end{aligned}$$

Periodic phenomena: a motivating example

Let $T^2 = (\mathbb{R}/\mathbb{Z})^2$ be the 2D torus. Consider the dynamical system given by

$$\begin{aligned} \Phi_\sigma : T^2 \times \mathbb{R} &\to T^2 \\ \big((a,b),t\big) &\mapsto (a+t,\ b+\sigma t) \end{aligned}$$

If σ is rational, then every orbit is periodic.



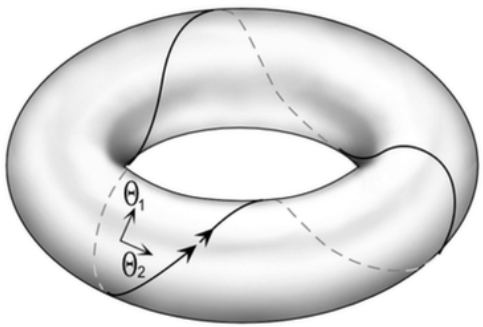
rational σ

Periodic phenomena: a motivating example

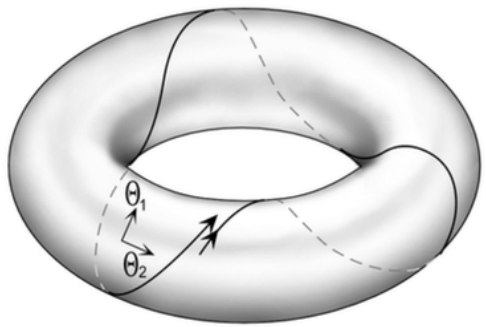
Let $T^2 = (\mathbb{R}/\mathbb{Z})^2$ be the 2D torus. Consider the dynamical system given by

$$\begin{aligned} \Phi_\sigma : T^2 \times \mathbb{R} &\to T^2 \\ \big((a,b),t\big) &\mapsto (a+t,\ b+\sigma t) \end{aligned}$$

If σ is rational, then every orbit is periodic. Otherwise every orbit is dense in T^2.



rational σ



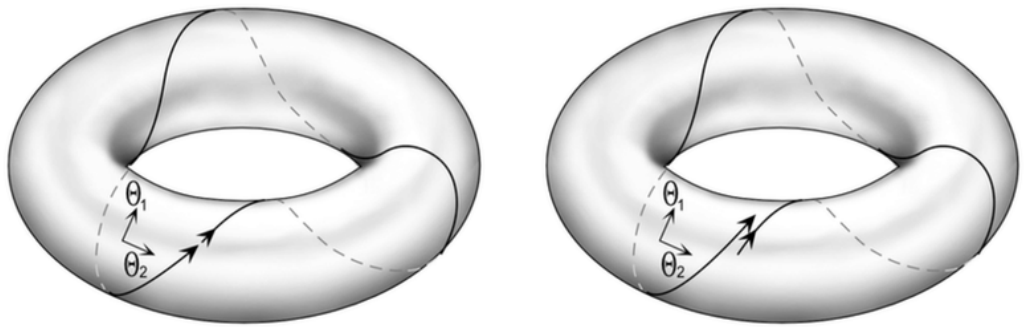
irrational σ

Periodic phenomena: a motivating example

Let $T^2 = (\mathbb{R}/\mathbb{Z})^2$ be the 2D torus. Consider the dynamical system given by

$$\begin{aligned} \Phi_\sigma : T^2 \times \mathbb{R} &\to T^2 \\ \big((a,b),t\big) &\mapsto (a+t,\ b+\sigma t) \end{aligned}$$

If σ is rational, then every orbit is periodic. Otherwise every orbit is dense in T^2.



rational σ *irrational* σ

From time series to topological shapes

Most periodic time series can be realized by a topological circle S^1 embedded in a Euclidean space of higher dimension.

Topological time series analysis

Let us make the assumption that sampled signals are distributed over a manifold (!) To topologically analyze time series, we then proceed as follows:

Topological time series analysis

Let us make the assumption that sampled signals are distributed over a manifold (!) To topologically analyze time series, we then proceed as follows:

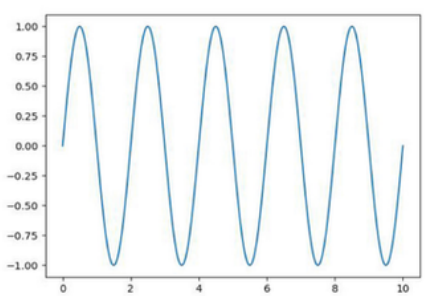
Step 1 Embed the data into a Euclidean space of suitable dimension;

Topological time series analysis

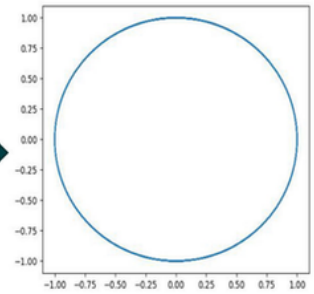
Let us make the assumption that sampled signals are distributed over a manifold (!) To topologically analyze time series, we then proceed as follows:

Step 1 Embed the data into a Euclidean space of suitable dimension;

Step 2 Compute the algebraic invariants for statistical inference.



realization



computation

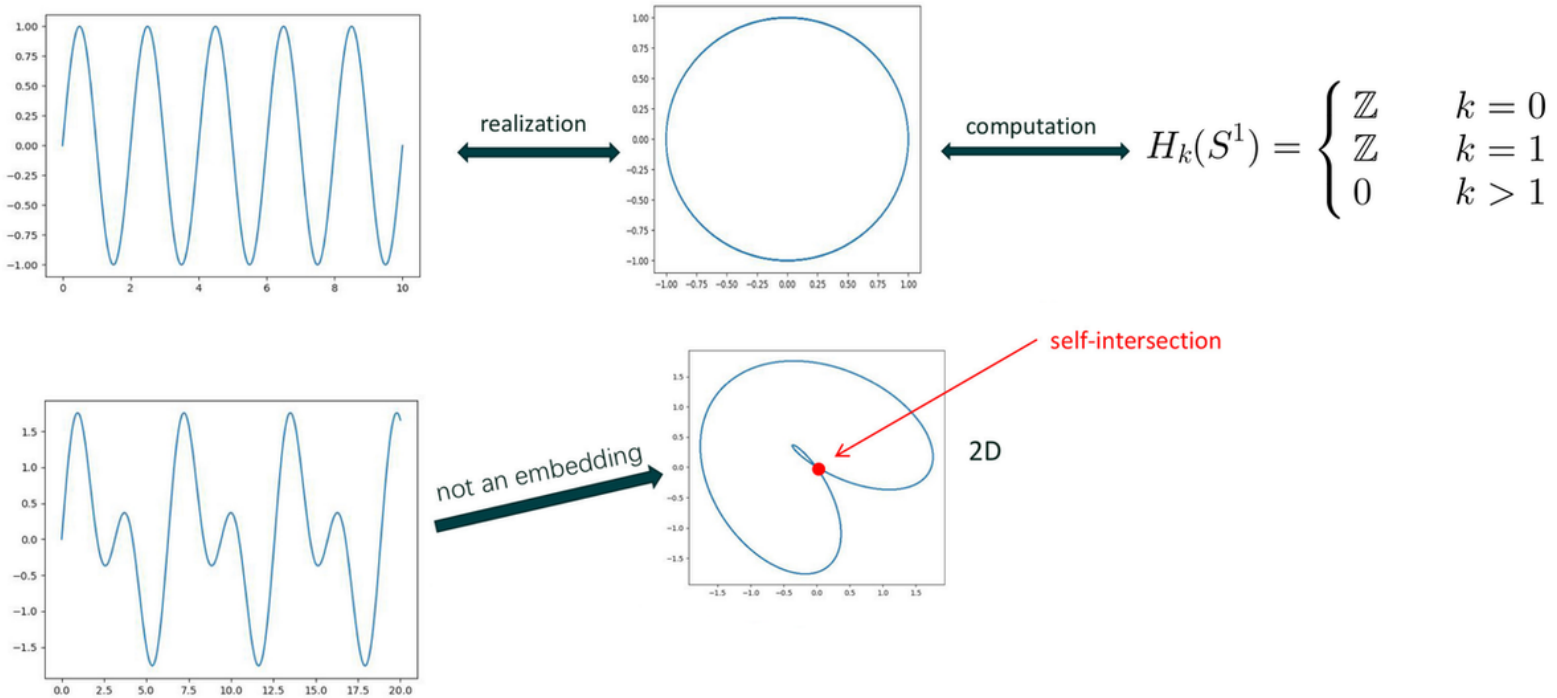
$$H_k(S^1) = \begin{cases} \mathbb{Z} & k = 0 \\ \mathbb{Z} & k = 1 \\ 0 & k > 1 \end{cases}$$

Topological time series analysis

Let us make the assumption that sampled signals are distributed over a manifold (!) To topologically analyze time series, we then proceed as follows:

Step 1 Embed the data into a Euclidean space of suitable dimension;

Step 2 Compute the algebraic invariants for statistical inference.

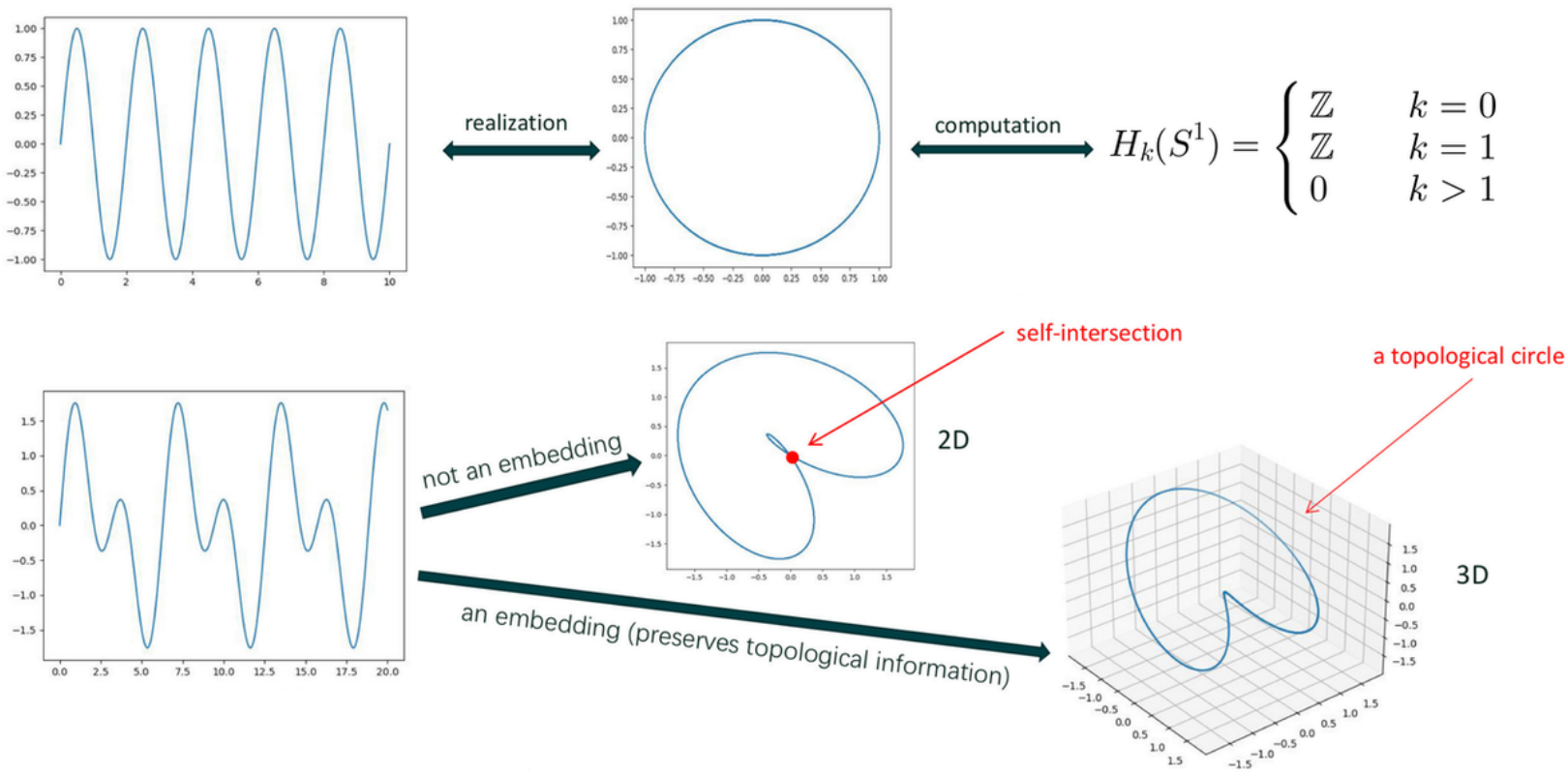


Topological time series analysis

Let us make the assumption that sampled signals are distributed over a manifold (!) To topologically analyze time series, we then proceed as follows:

Step 1 Embed the data into a Euclidean space of suitable dimension;

Step 2 Compute the algebraic invariants for statistical inference.



Topological time series analysis

Let us make the assumption that sampled signals are distributed over a manifold (!) To topologically analyze time series, we then proceed as follows:

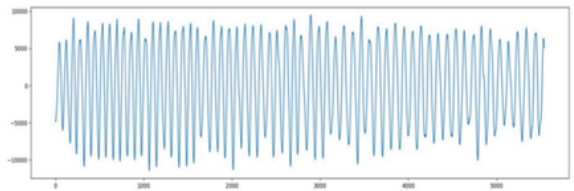
Step 1 Embed the data into a Euclidean space of suitable dimension;

Step 2 Compute the algebraic invariants for statistical inference.

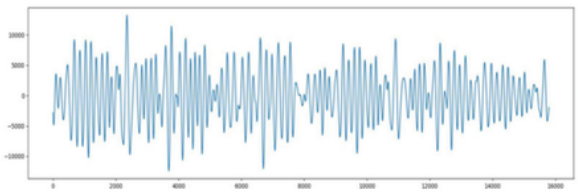
- Perea, *Topological time series analysis*, **Notices of the American Mathematical Society**, 2019.
- Perea & Harer, *Sliding windows and persistence: An application of topological methods to signal analysis*, **Foundations of Computational Mathematics**, 2015.

An application: detection of wheeze in medical science (pulmonology)

wheeze

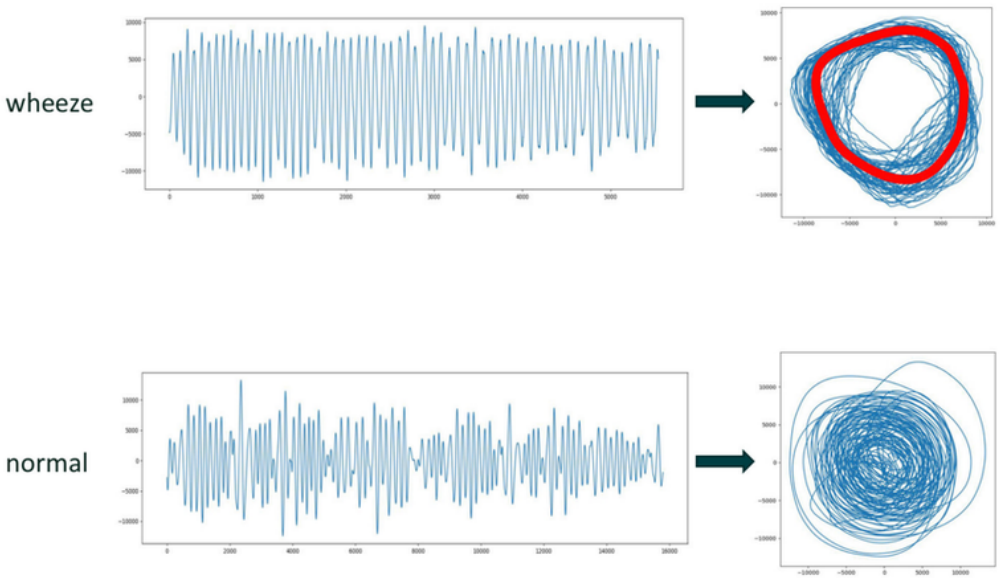


normal



Original sound signals

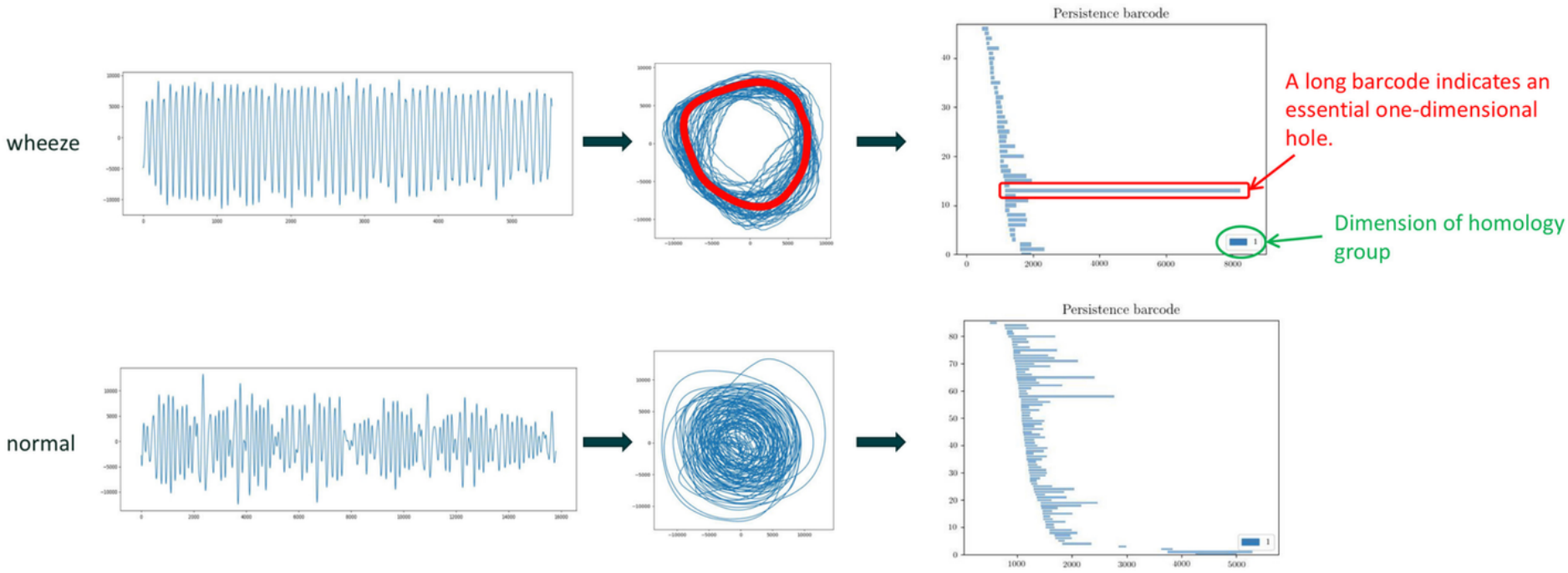
An application: detection of wheeze in medical science (pulmonology)



Original sound signals

Realized topological shapes embedded in 2D Euclidean space

An application: detection of wheeze in medical science (pulmonology)



Original sound signals

Realized topological shapes embedded in 2D Euclidean space

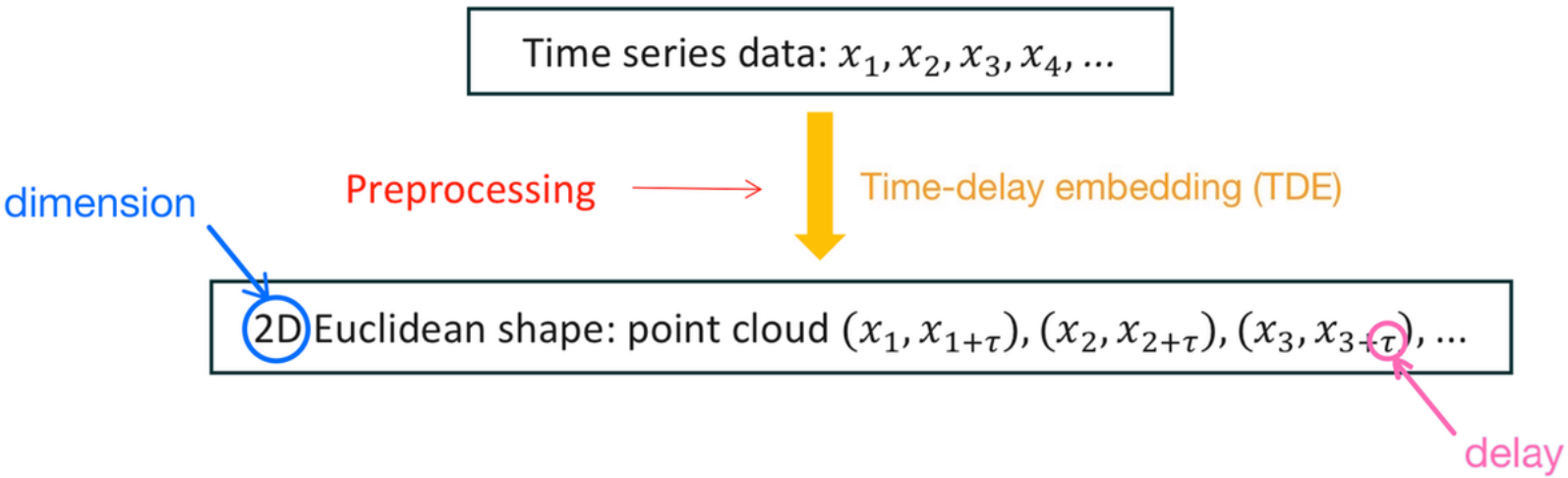
Persistence barcodes as representations of the algebraic invariant (1D homology group)

Emrani, Gentimis & Krim, *Persistent homology of delay embeddings and its application to wheeze detection*, **IEEE Signal Processing Letters**, 2014.

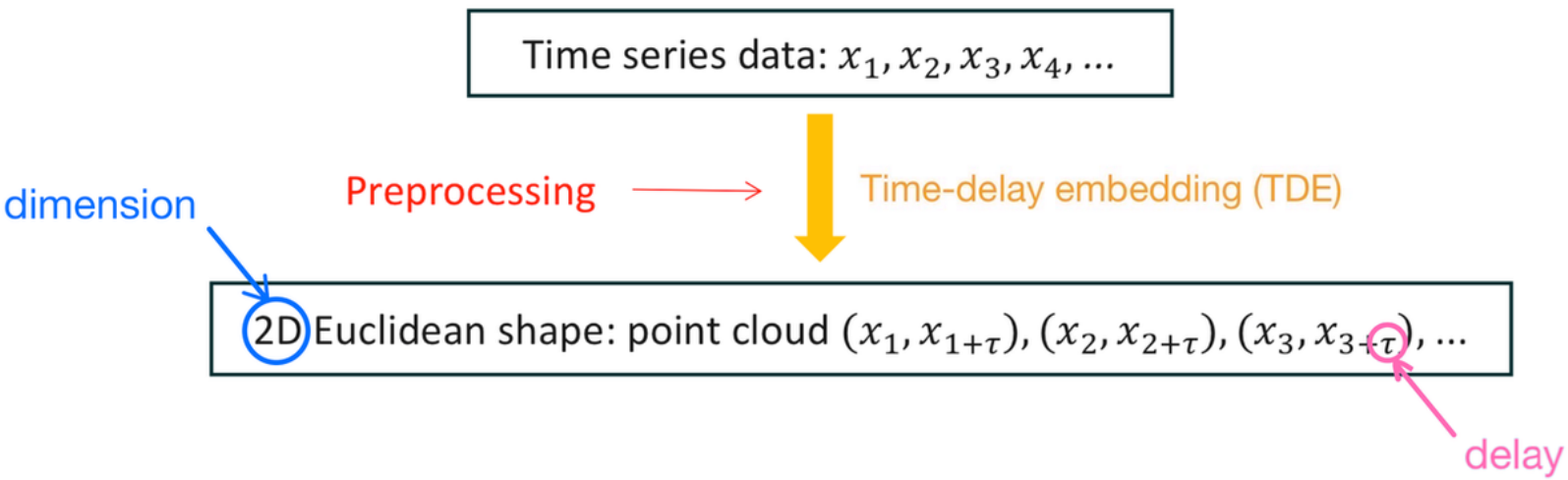
A pipeline for topological time series analysis

Time series data: $x_1, x_2, x_3, x_4, \ldots$

A pipeline for topological time series analysis



A pipeline for topological time series analysis



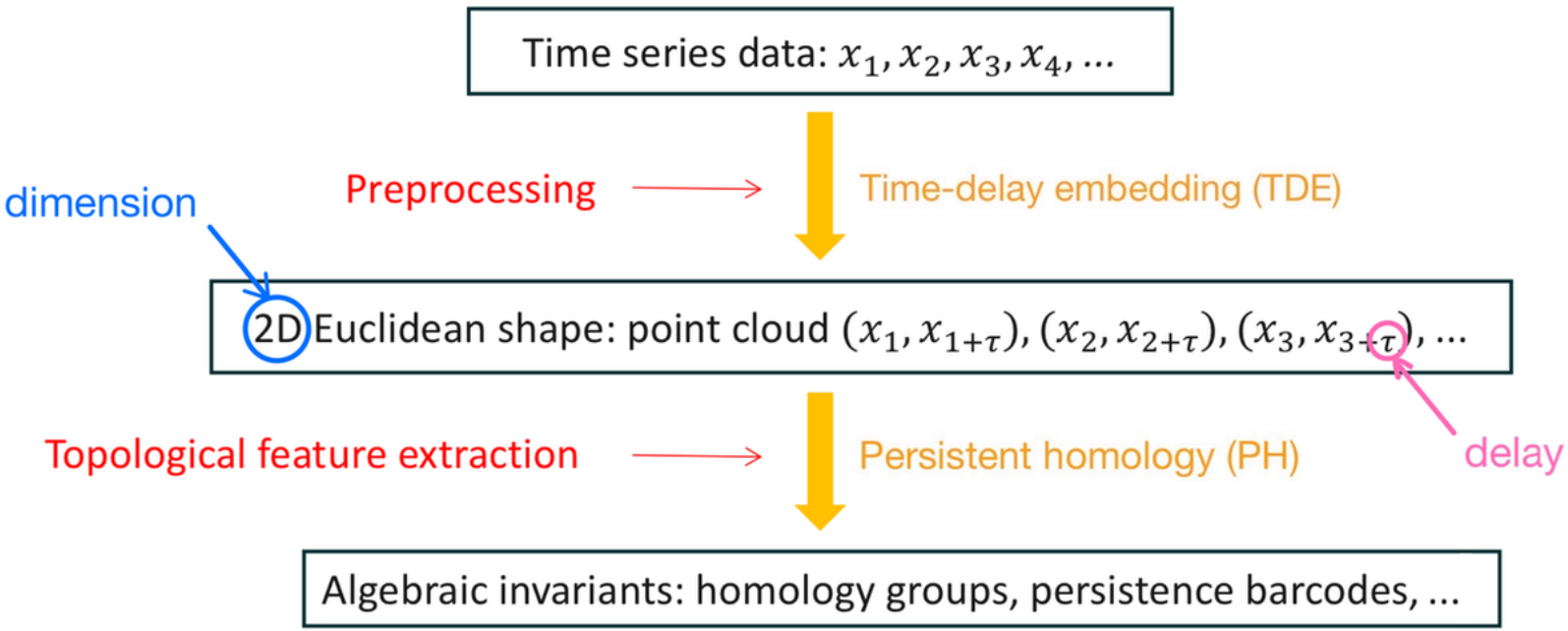
Euclidean embedding of time series data dates back to Takens's work on fluid turbulence.

Theorem (Takens 1981). Let M be a compact manifold of dimension n. Given pairs (φ, y) with $\varphi: M \to M$ a smooth diffeomorphism and $y: M \to \mathbb{R}$ a smooth function, it is a generic property that the map $\Phi_{(\varphi,y)}: M \to \mathbb{R}^{2n+1}$ defined by

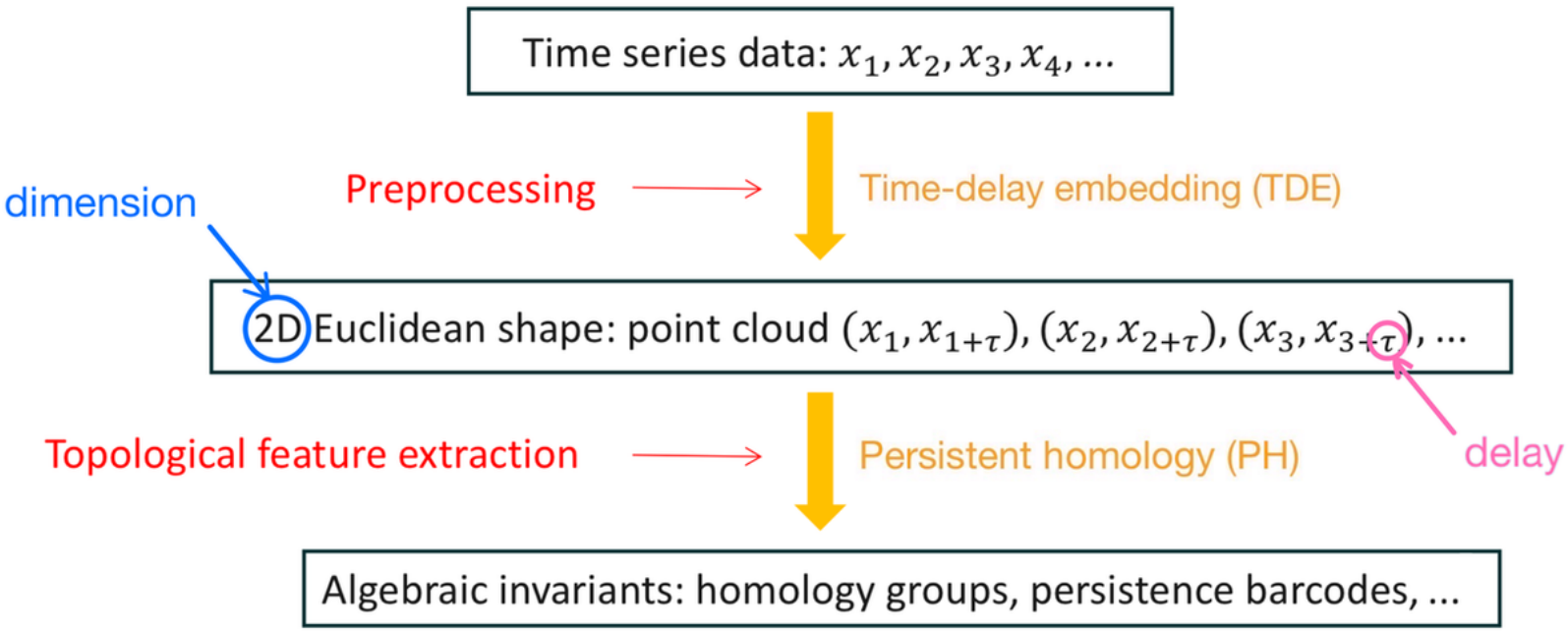
$$\Phi_{(\varphi,y)}(x) = \left(y(x), y(\varphi(x)), \ldots, y(\varphi^{2n}(x))\right)$$

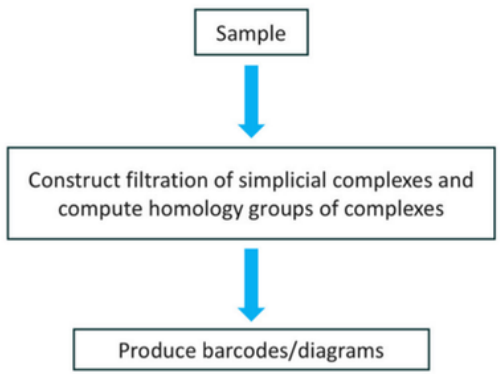
is an embedding.

A pipeline for topological time series analysis

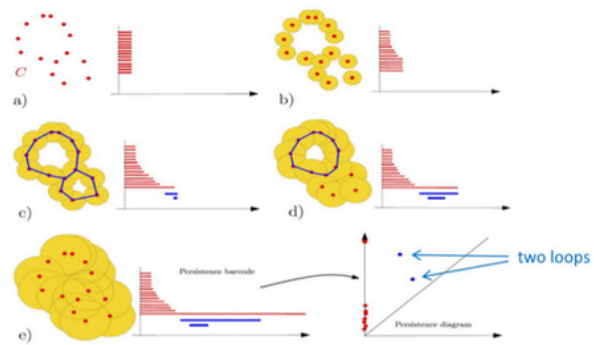


A pipeline for topological time series analysis

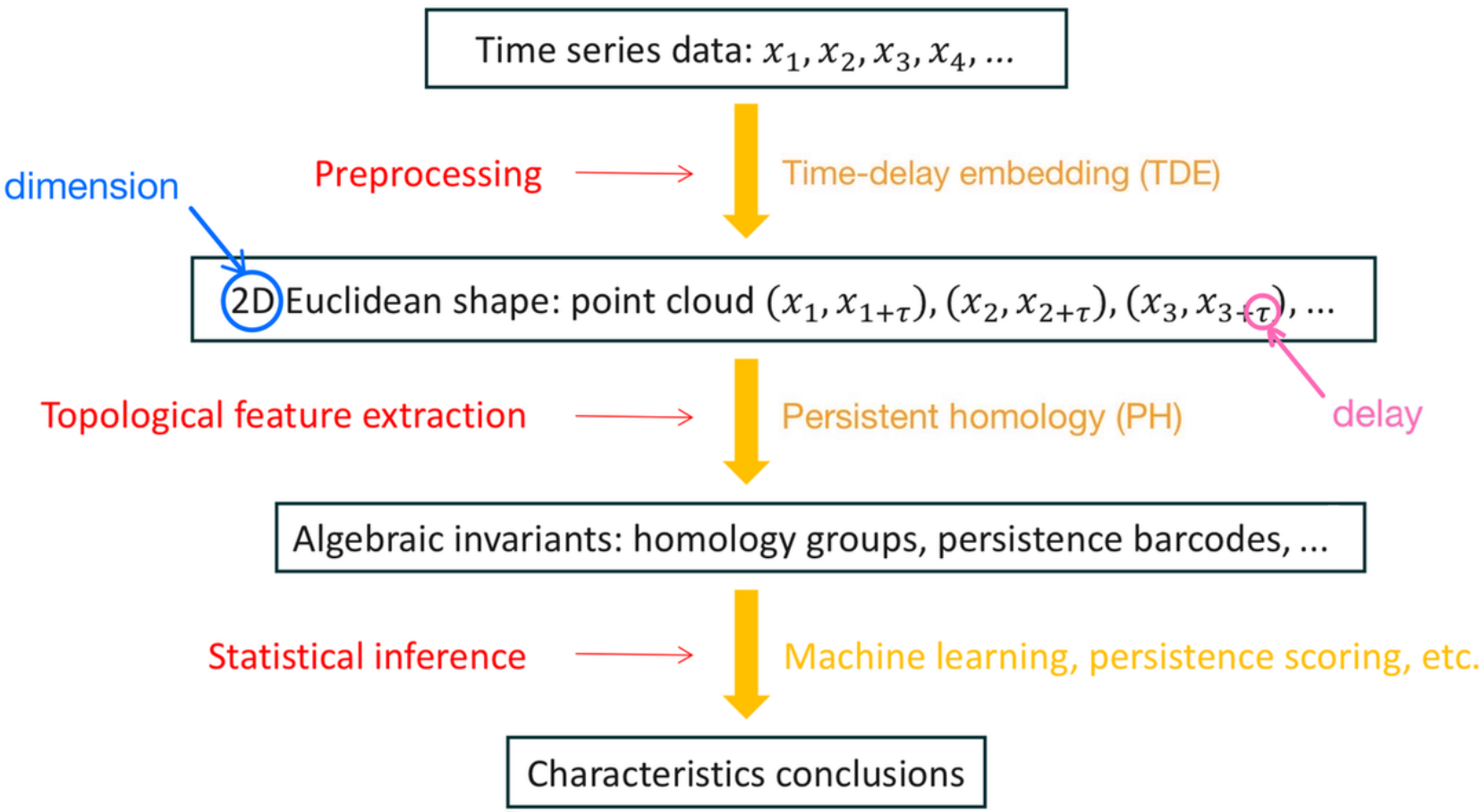




How filtration through varying distance measure reveals essential topological features



A pipeline for topological time series analysis



Classification of speech signals

In consultation with Meng Yu of Tencent AI Lab, we applied topological methods to classify voiced/voiceless and vowel/consonant speech data

Classification of speech signals

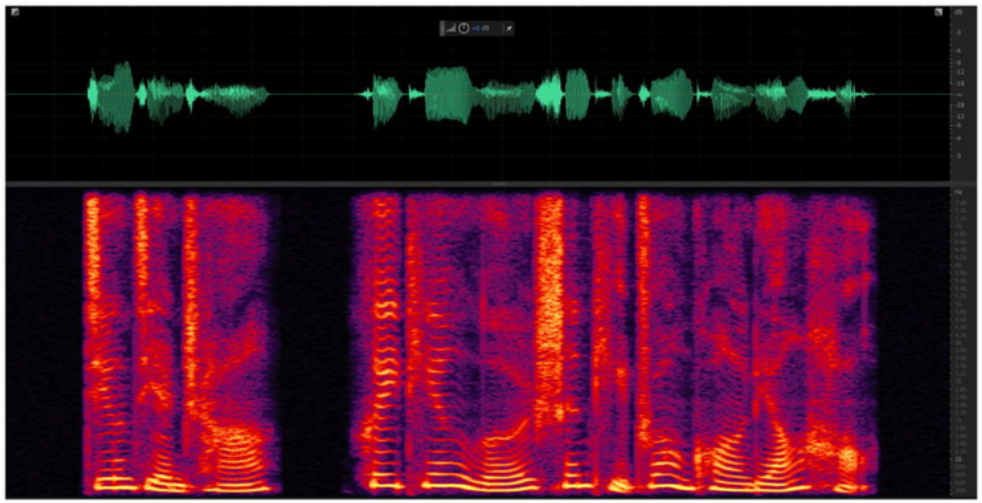
In consultation with Meng Yu of Tencent AI Lab, we applied topological methods to classify voiced/voiceless and vowel/consonant speech data, with motivations from industrial applications, including medical diagnosis, neurophysiology, speaker identification and other AI innovations.

Classification of speech signals

In consultation with Meng Yu of Tencent AI Lab, we applied topological methods to classify voiced/voiceless and vowel/consonant speech data, with motivations from industrial applications, including medical diagnosis, neurophysiology, speaker identification and other AI innovations.

Spectrograms

There are speech signal processing softwares for professional use.

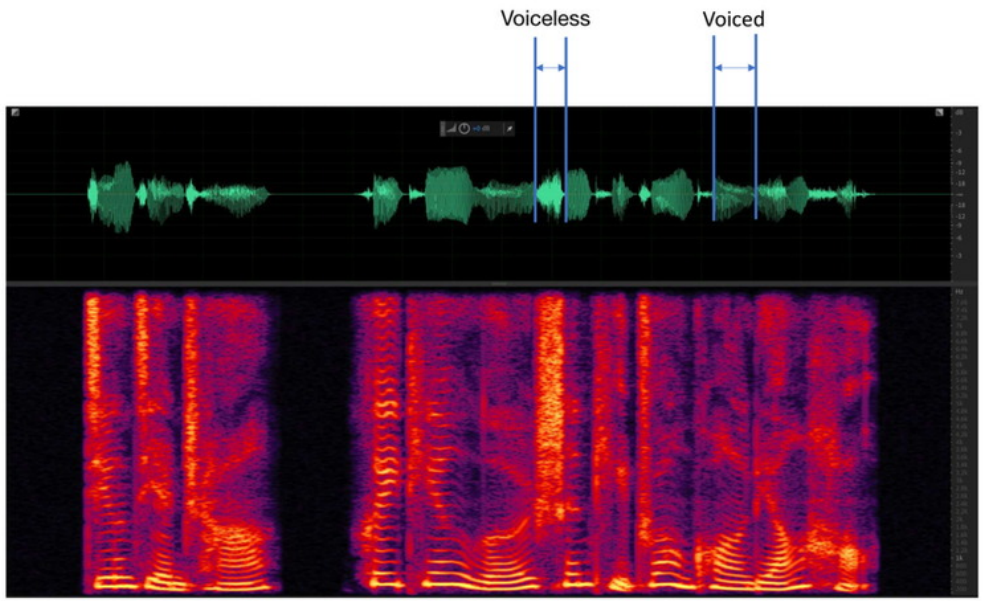


Classification of speech signals

In consultation with Meng Yu of Tencent AI Lab, we applied topological methods to classify voiced/voiceless and vowel/consonant speech data, with motivations from industrial applications, including medical diagnosis, neurophysiology, speaker identification and other AI innovations.

Spectrograms

There are speech signal processing softwares for professional use.



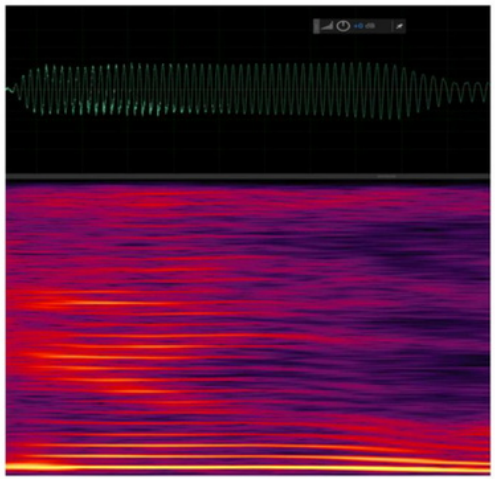
Classification of speech signals

Voiced

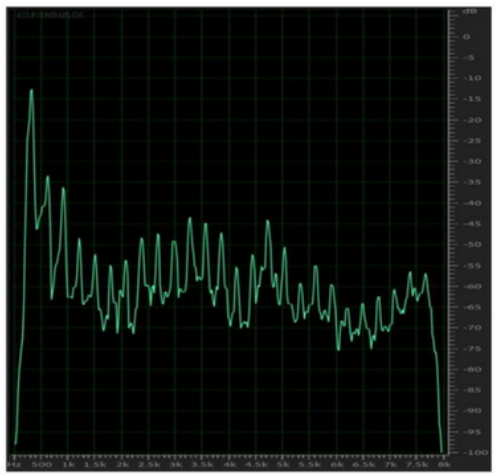
[ŋ], [m], [n], [j], [l], [v], [ʒ], *etc.*

Sinusoid in time domain

Harmonics in frequency domain



Time and Time-Frequency domain

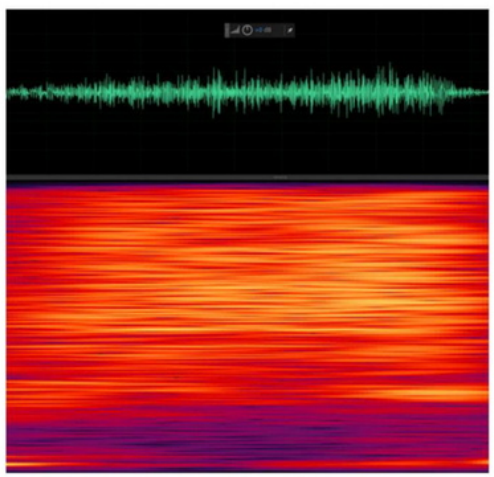


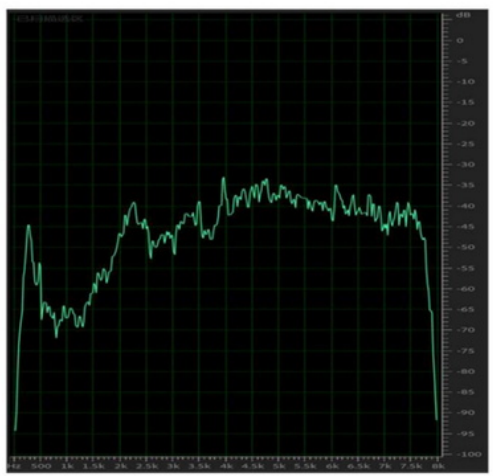
Frequency response

Like a white noise

Voiceless

[f], [k], [θ], [t], [s], [tʃ], *etc.*





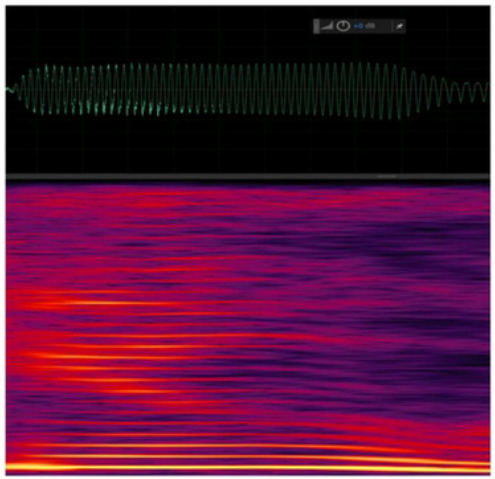
Classification of speech signals

Voiced

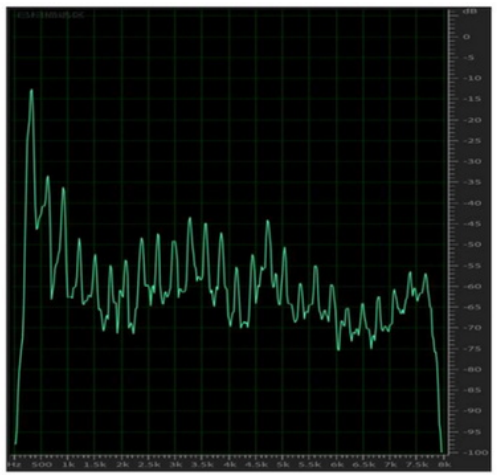
exhibit periodic waveforms resulting from glottal vibrations

Sinusoid in time domain

Harmonics in frequency domain



Time and Time-Frequency domain

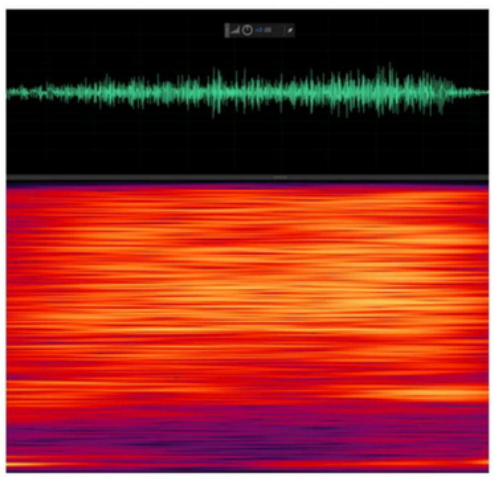


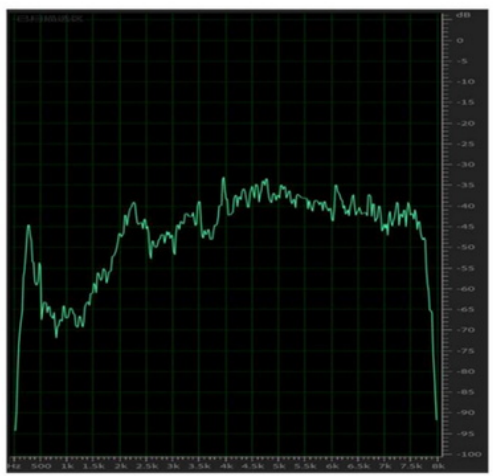
Frequency response

Voiceless

predominantly characterized by aperiodic, turbulence-induced noise

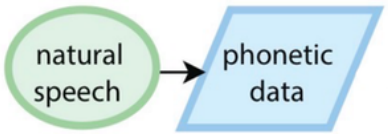
Like a white noise





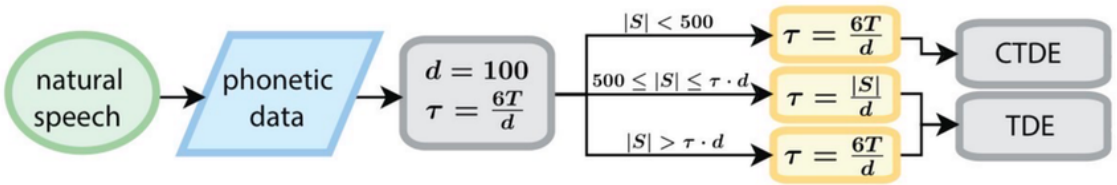
Primary experiments combining topological features with ML models

Here is a flowchart for our method of TopCap:



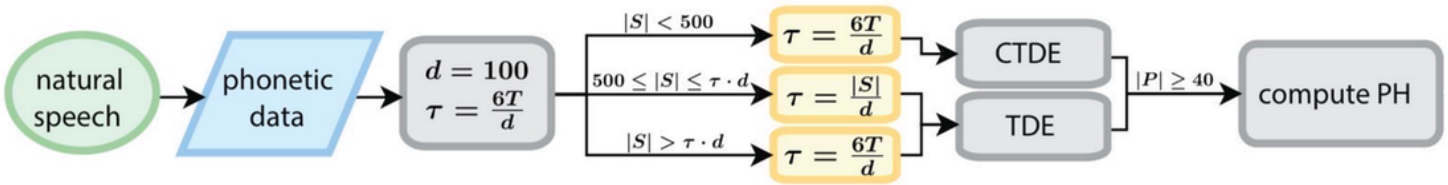
Primary experiments combining topological features with ML models

Here is a flowchart for our method of TopCap:



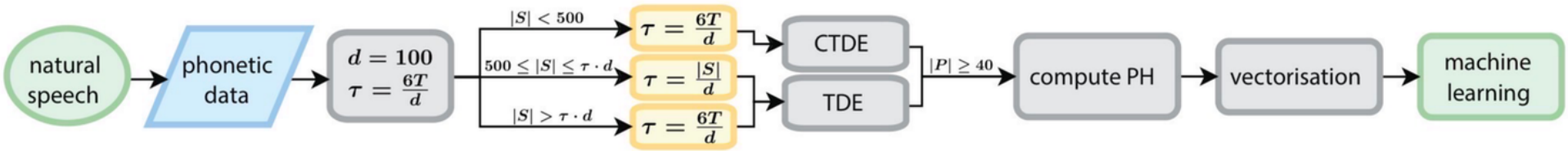
Primary experiments combining topological features with ML models

Here is a flowchart for our method of TopCap:



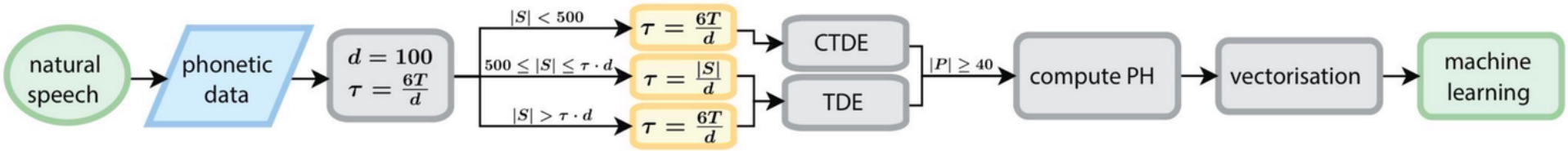
Primary experiments combining topological features with ML models

Here is a flowchart for our method of TopCap:

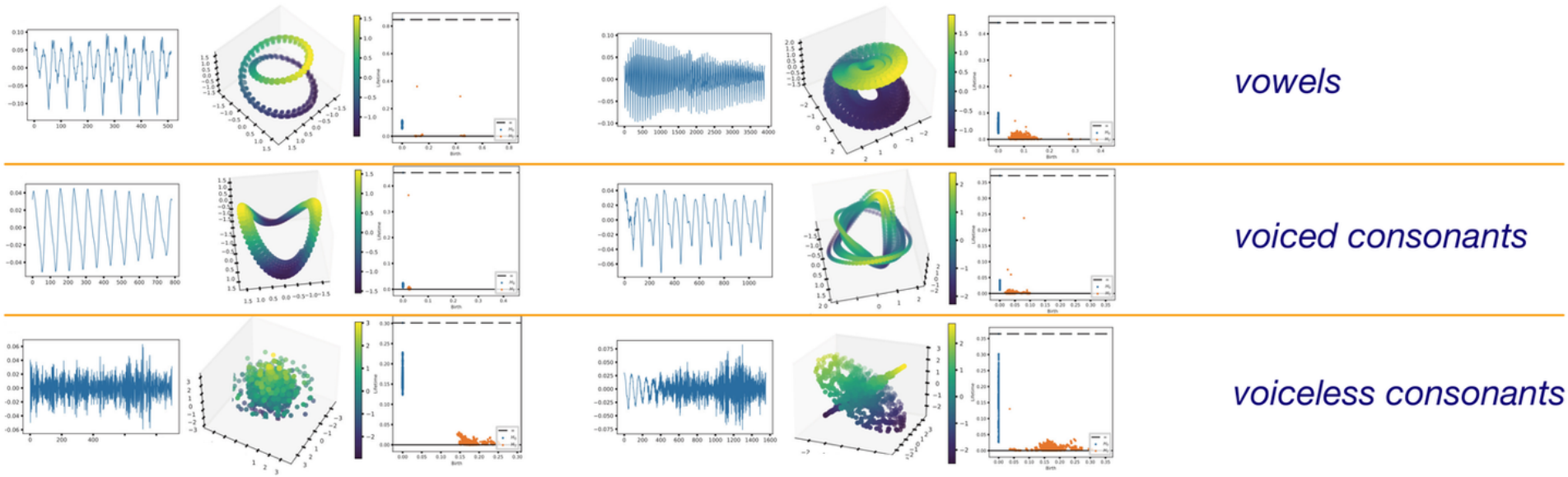


Primary experiments combining topological features with ML models

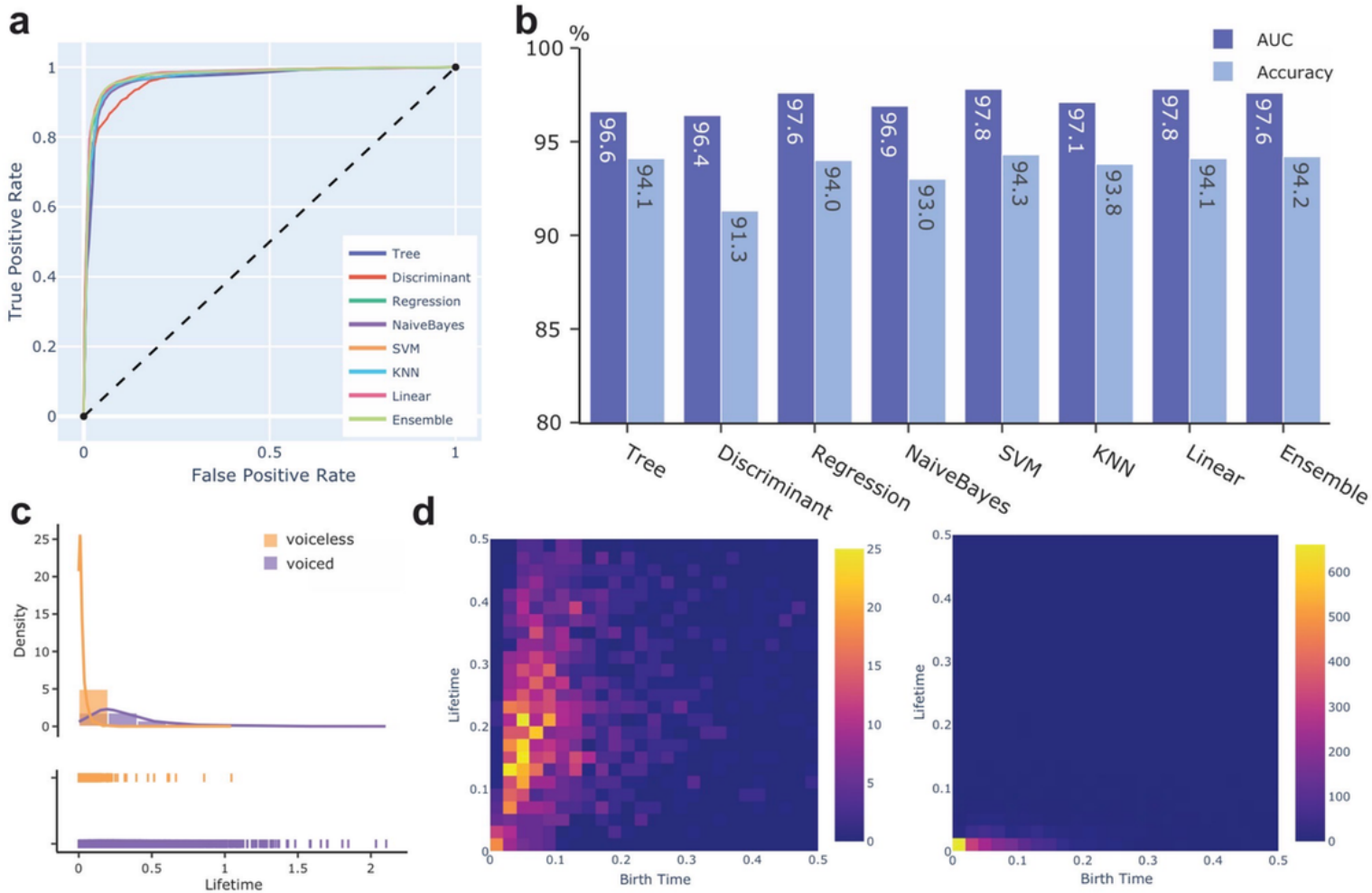
Here is a flowchart for our method of TopCap:



Topological profiles for vowels and consonants

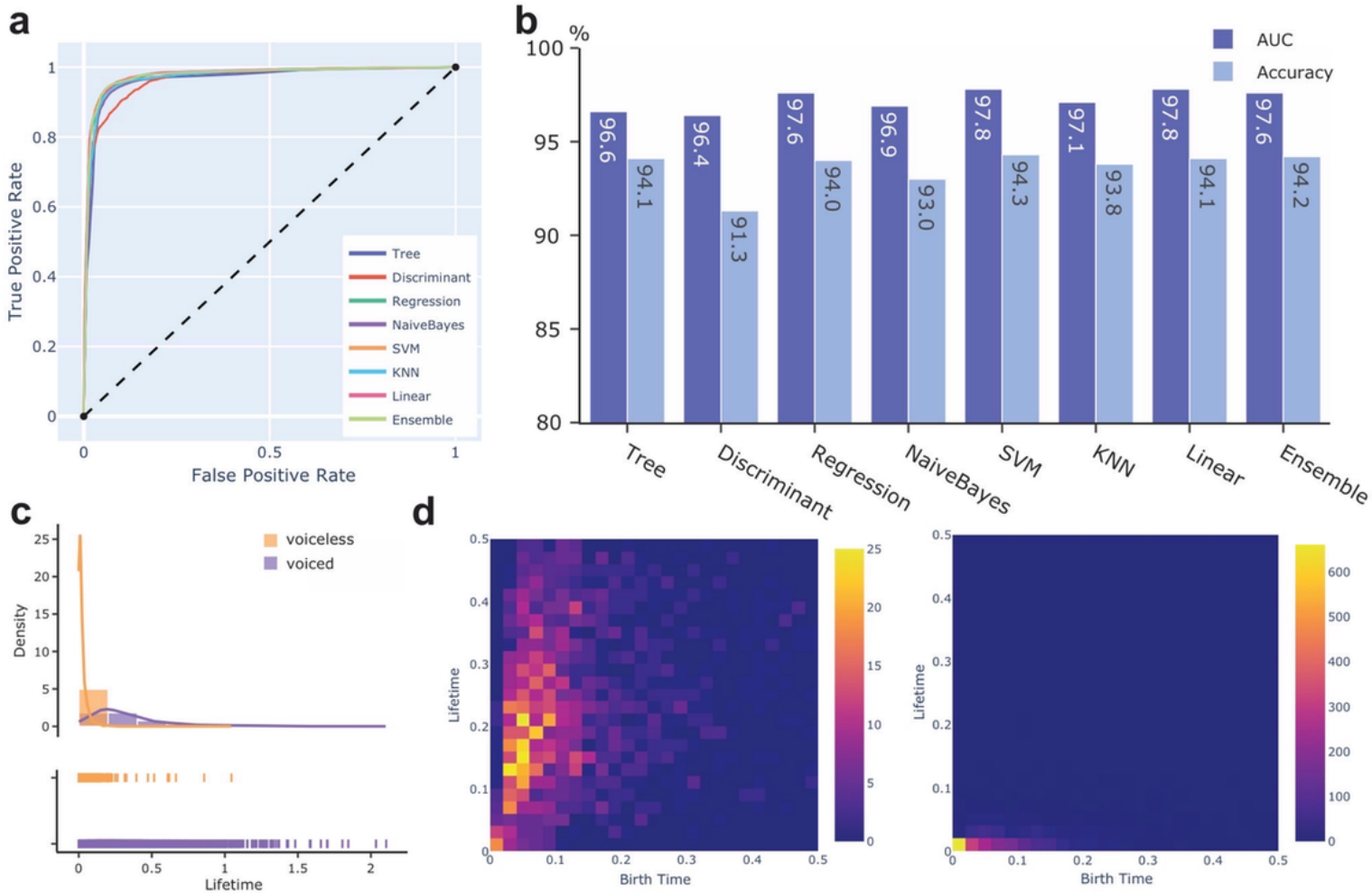


Primary experiments combining topological features with ML models



Results of machine learning topological features

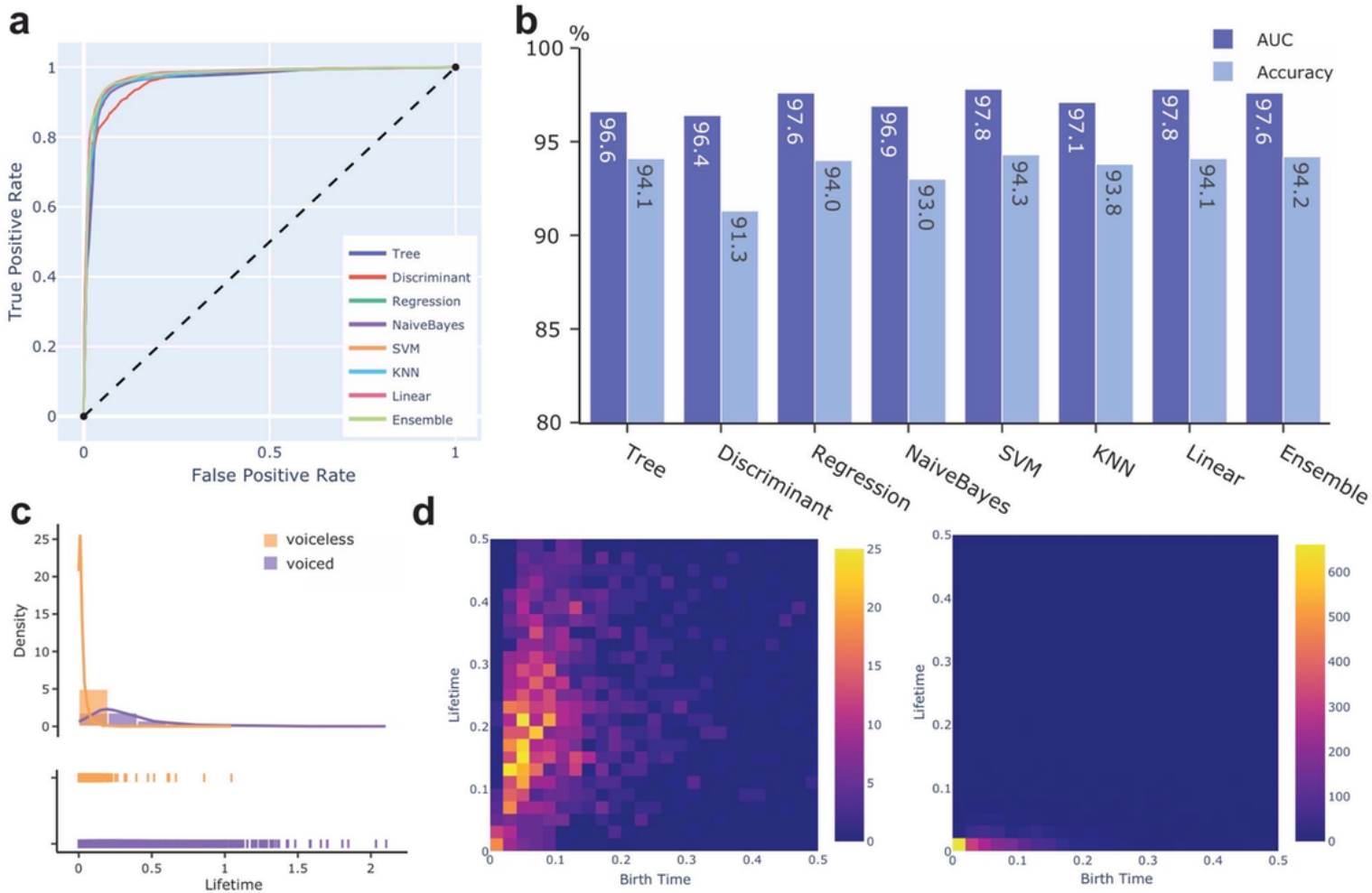
Primary experiments combining topological features with ML models



Results of machine learning topological features

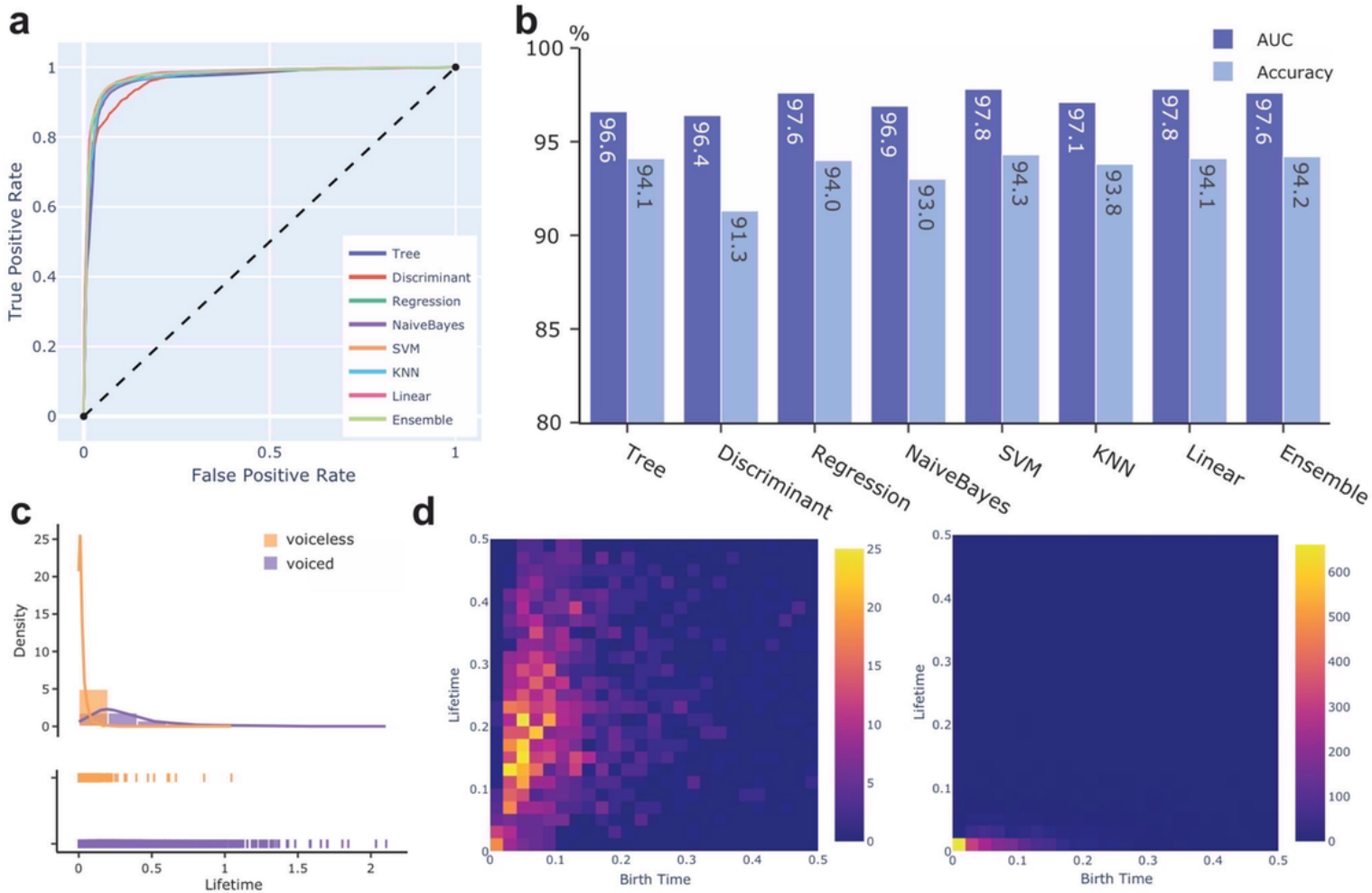
***a**, Receiver operating characteristic curves of traditional machine learning algorithms.*

Primary experiments combining topological features with ML models



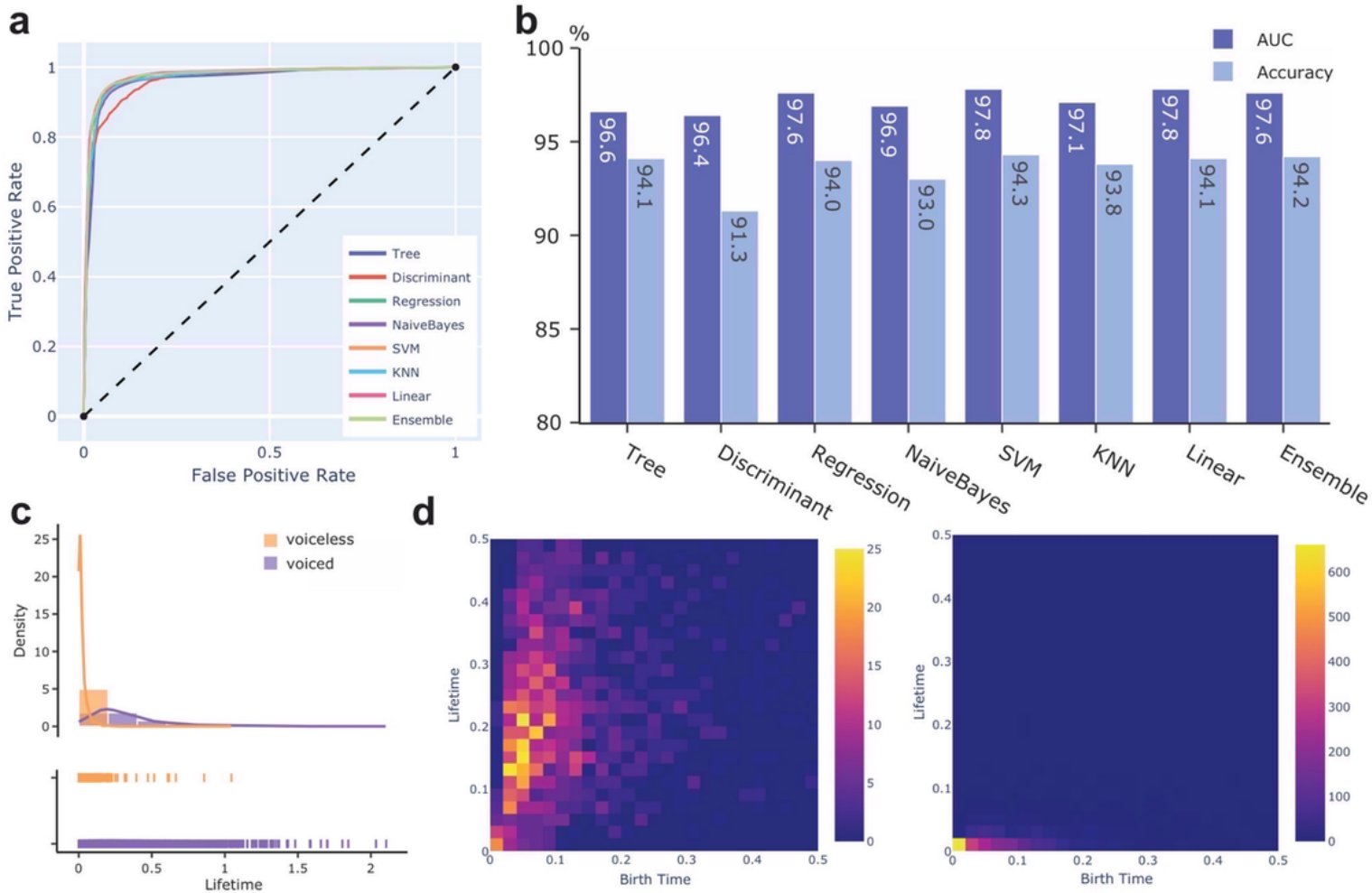
Results of machine learning topological features
***b**, Accuracy and area under the curve of each of these algorithms.*

Primary experiments combining topological features with ML models



Results of machine learning topological features
***c**, Histograms of records represented by their PH-lifetime for voiced and voiceless consonants, together with kernel density estimation and rug plot. The distributions of maximal persistence can distinguish voiced and voiceless consonants.*

Primary experiments combining topological features with ML models



Results of machine learning topological features
***d**, Diagrams of records represented as (birth time, lifetime) for voiced consonants (left) and voiceless consonants (right), where voiced consonants exhibit higher birth time and lifetime. The color represents the density of points in each unit grid box.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Accuracy rates % of TopCap on 8 small datasets and 4 large datasets stand in comparison with state-of-the-art methods.

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Accuracy rates % of TopCap on 8 small datasets and 4 large datasets stand in comparison with state-of-the-art methods. While MFCC–Transformer and STFT–CNN-16 generally outperform TopCap, it is important to note that TopCap exceeds the performance of MFCC–GRU (gated recurrent unit, which also uses advanced architecture) and STFT–CNN-8 (convolutional neural network, a smaller model than STFT–CNN-16) on small datasets.

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Accuracy rates % of TopCap on 8 small datasets and 4 large datasets stand in comparison with state-of-the-art methods. While MFCC–Transformer and STFT–CNN-16 generally outperform TopCap, it is important to note that TopCap exceeds the performance of MFCC–GRU (gated recurrent unit, which also uses advanced architecture) and STFT–CNN-8 (convolutional neural network, a smaller model than STFT–CNN-16) on small datasets. For larger datasets, TopCap generally does not match the performance of deep neural networks, primarily due to its use of simpler topological features and basic machine learning models.

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Accuracy rates % of TopCap on 8 small datasets and 4 large datasets stand in comparison with state-of-the-art methods. While MFCC–Transformer and STFT–CNN-16 generally outperform TopCap, it is important to note that TopCap exceeds the performance of MFCC–GRU (gated recurrent unit, which also uses advanced architecture) and STFT–CNN-8 (convolutional neural network, a smaller model than STFT–CNN-16) on small datasets. For larger datasets, TopCap generally does not match the performance of deep neural networks, primarily due to its use of simpler topological features and basic machine learning models. This limitation motivates the integration TopNN of topological features into neural networks.

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Accuracy rates % of TopCap on 8 small datasets and 4 large datasets stand in comparison with state-of-the-art methods. While MFCC–Transformer and STFT–CNN-16 generally outperform TopCap, it is important to note that TopCap exceeds the performance of MFCC–GRU (gated recurrent unit, which also uses advanced architecture) and STFT–CNN-8 (convolutional neural network, a smaller model than STFT–CNN-16) on small datasets. For larger datasets, TopCap generally does not match the performance of deep neural networks, primarily due to its use of simpler topological features and basic machine learning models. This limitation motivates the integration TopNN of topological features into neural networks. Overall, while TopCap may not achieve the highest performance across all benchmarks, it produces decent results.

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Structural efficiency.*** *Neural network models require further feature extraction from input MFCC sequences or STFT spectrograms for classification tasks, necessitating a training process which lengthens with a growing dataset.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Structural efficiency.*** *Neural network models require further feature extraction from input MFCC sequences or STFT spectrograms for classification tasks, necessitating a training process which lengthens with a growing dataset. In contrast, TopCap mainly utilizes topology-based methods (TDE and PH) which are more straightforward for feature extraction.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Structural efficiency.*** *Neural network models require further feature extraction from input MFCC sequences or STFT spectrograms for classification tasks, necessitating a training process which lengthens with a growing dataset. In contrast, TopCap mainly utilizes topology-based methods (TDE and PH) which are more straightforward for feature extraction. Meanwhile, the topological fingerprints (e.g., maximal persistence) are strong enough to characterize phonemes effectively for our classification tasks.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Structural efficiency.*** *Neural network models require further feature extraction from input MFCC sequences or STFT spectrograms for classification tasks, necessitating a training process which lengthens with a growing dataset. In contrast, TopCap mainly utilizes topology-based methods (TDE and PH) which are more straightforward for feature extraction. Meanwhile, the topological fingerprints (e.g., maximal persistence) are strong enough to characterize phonemes effectively for our classification tasks. Therefore, TopCap gains higher efficiency, especially when handling larger datasets.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Structural efficiency.*** *Neural network models require further feature extraction from input MFCC sequences or STFT spectrograms for classification tasks, necessitating a training process which lengthens with a growing dataset. In contrast, TopCap mainly utilizes topology-based methods (TDE and PH) which are more straightforward for feature extraction. Meanwhile, the topological fingerprints (e.g., maximal persistence) are strong enough to characterize phonemes effectively for our classification tasks. Therefore, TopCap gains higher efficiency, especially when handling larger datasets. On a related note, deep learning methods, as a data-driven approach, require large amounts of data for training and generalization.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Structural efficiency.*** *Neural network models require further feature extraction from input MFCC sequences or STFT spectrograms for classification tasks, necessitating a training process which lengthens with a growing dataset. In contrast, TopCap mainly utilizes topology-based methods (TDE and PH) which are more straightforward for feature extraction. Meanwhile, the topological fingerprints (e.g., maximal persistence) are strong enough to characterize phonemes effectively for our classification tasks. Therefore, TopCap gains higher efficiency, especially when handling larger datasets. On a related note, deep learning methods, as a data-driven approach, require large amounts of data for training and generalization. In contrast, comparing the upper and lower halves of the above table, we see that TopCap achieves equally good performance on relatively small datasets.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Interpretability.*** *Neural networks are often referred to as black boxes due to their low explainability and interpretability, which makes it challenging to understand the mechanism of feature extraction and effectively improve a model for classification. However, TopCap offers a white-box method for visualizing features of time series data, which gives insight to the intrinsic properties and nuanced differences within the data, enabling us to better understand and improve the model.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Computational speed.*** *Neural networks involve time-consuming training process, even with GPU acceleration. For instance, on the TIMIT dataset, a full training cycle of 15 epochs can take approximately 30 minutes with GPU parallelization.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

Advantages of TopCap:

- ***Computational speed.*** *Neural networks involve time-consuming training process, even with GPU acceleration. For instance, on the TIMIT dataset, a full training cycle of 15 epochs can take approximately 30 minutes with GPU parallelization. In contrast, TopCap bypasses the need for iterative training and achieves significantly faster computation. TopCap performs lightweight machine learning with negligible runtime overhead, completing both feature extraction and classification in just 2 minutes when utilizing 16-thread CPU parallelization.*

Model comparison on benchmark datasets

	ALLSSTAR corpora					Random samples		
Small dataset	HT1	HT2	DHR	LPP	NWS	LJ	TIMIT	Libri
Number of phones	3200	3000	3600	3800	1800	2000	2000	2000
TopCap	94.3	92.7	92.3	91.9	88.8	94.6	83.9	85.1
MFCC–GRU	93.3	92.2	93.2	91.4	89.8	86.0	70.5	79.0
MFCC–Transformer	96.0	93.9	94.2	92.4	94.4	92.0	96.3	87.5
STFT–CNN-8	87.1	84.0	78.2	79.1	79.9	82.7	76.3	77.5
STFT–CNN-16	96.7	95.1	94.4	92.1	94.0	95.6	89.4	88.7

Large dataset	ALLSSTAR	LJSpeech	TIMIT	LibriSpeech
Number of phones	21000	257000	42000	500000
TopCap	92.5	92.9	92.8	88.7
MFCC–GRU	93.9	96.2	97.4	91.0
MFCC–Transformer	93.7	96.9	97.6	92.1
STFT–CNN-8	81.2	85.4	77.5	80.3
STFT–CNN-16	94.6	96.3	91.4	90.6

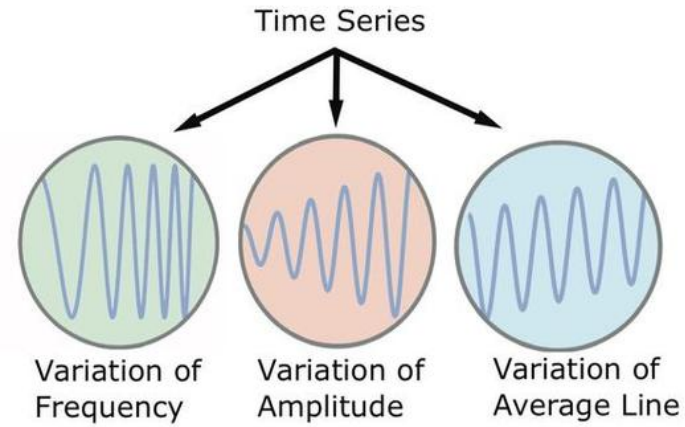
Advantages of TopCap:

- ***Computational speed.*** *Neural networks involve time-consuming training process, even with GPU acceleration. For instance, on the TIMIT dataset, a full training cycle of 15 epochs can take approximately 30 minutes with GPU parallelization. In contrast, TopCap bypasses the need for iterative training and achieves significantly faster computation. TopCap performs lightweight machine learning with negligible runtime overhead, completing both feature extraction and classification in just 2 minutes when utilizing 16-thread CPU parallelization. TopCap's efficiency advantage comes from avoiding gradient-based optimization and using computationally cheaper topologically derived features, along with a highly parallelizable pipeline. These make it significantly faster and more scalable especially for large datasets or real-time applications.*

Detection of vibration patterns

TopCap descriptors probe finer information such as the vibration of a time series.

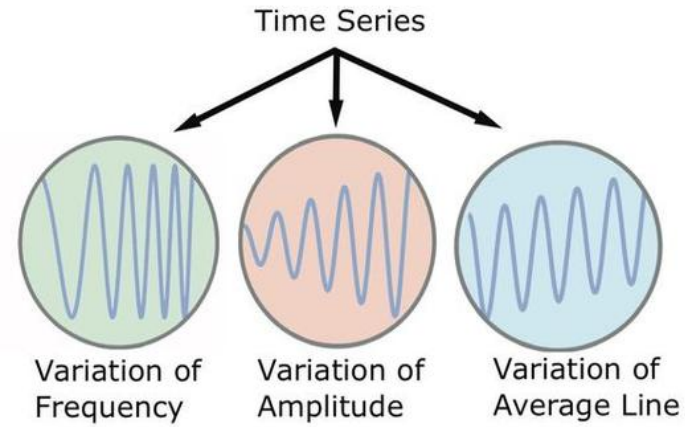
- Three fundamental variations: frequency, amplitude, average line.

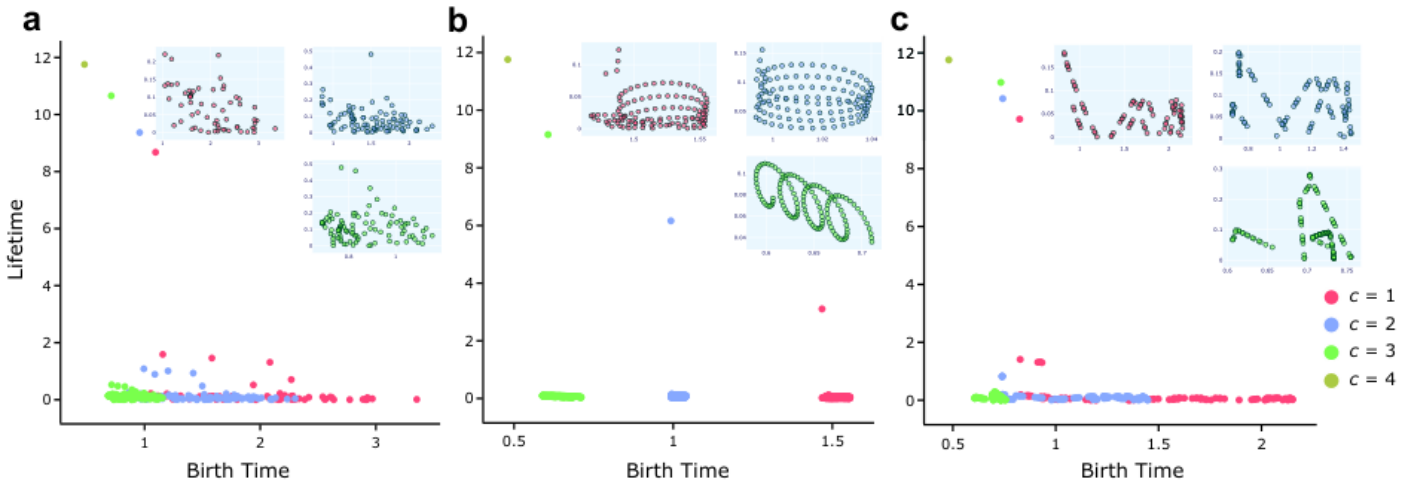


Detection of vibration patterns

TopCap descriptors probe finer information such as the vibration of a time series.

- Three fundamental variations: frequency, amplitude, average line.
- Points near the birth–death diagonal distinguish these variations.
- To investigate: topological featurization for tonal languages according to variation of frequency.





1-dimensional PH distinguishes the three fundamental variations.

From topological data analysis to topological deep learning

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces

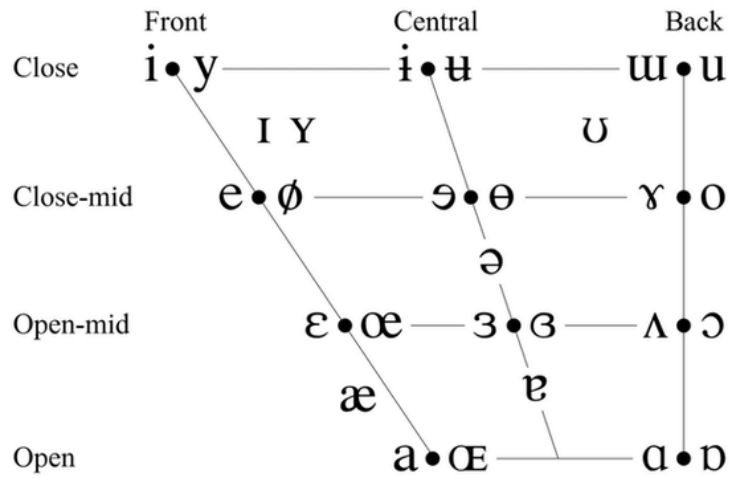
From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

• For phonetic data, linguists created a charted “distribution space” of vowels:

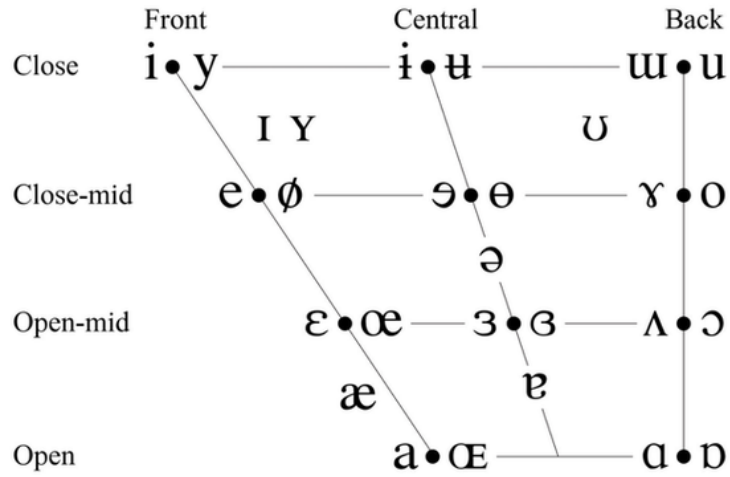


From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

• For phonetic data, linguists created a charted “distribution space” of vowels:

The vertical axis of the chart denotes vowel height. Vowels pronounced with the tongue lowered are at the bottom and raised are at the top.

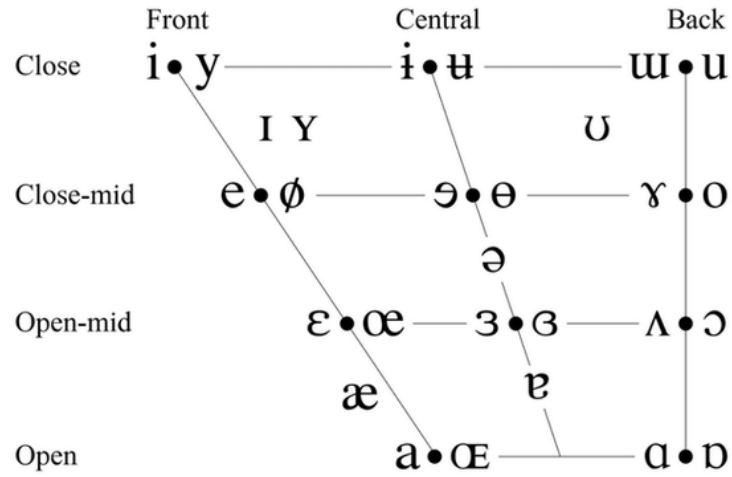


From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

• For phonetic data, linguists created a charted “distribution space” of vowels:

The vertical axis of the chart denotes vowel height. Vowels pronounced with the tongue lowered are at the bottom and raised are at the top. The horizontal axis of the chart denotes vowel backness. Vowels with the tongue moved towards the front of the mouth are in the left of the chart, while those with the tongue moved to the back are I placed in right.

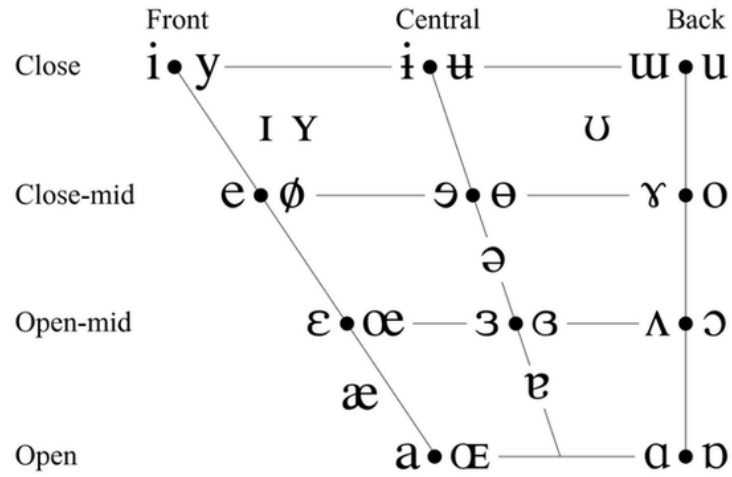


From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

• For phonetic data, linguists created a charted “distribution space” of vowels:

The vertical axis of the chart denotes vowel height. Vowels pronounced with the tongue lowered are at the bottom and raised are at the top. The horizontal axis of the chart denotes vowel backness. Vowels with the tongue moved towards the front of the mouth are in the left of the chart, while those with the tongue moved to the back are I placed in right. The last parameter is whether the lips are rounded. At each given spot, vowels on the right and left are rounded and unrounded, respectively.



From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

- For phonetic data, linguists created a charted “distribution space” of vowels.

- A main goal remains to use topological methods to reveal a distribution space for speech (and audio) data

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

- For phonetic data, linguists created a charted “distribution space” of vowels.

- A main goal remains to use topological methods to reveal a distribution space for speech (and audio) data, even a digraph on it modeling the complex network of speech-signal sequences

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

- For phonetic data, linguists created a charted “distribution space” of vowels.

- A main goal remains to use topological methods to reveal a distribution space for speech (and audio) data, even a digraph on it modeling the complex network of speech-signal sequences, and apply these topological inputs for smarter learning.

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

- For phonetic data, linguists created a charted “distribution space” of vowels.

- A main goal remains to use topological methods to reveal a distribution space for speech (and audio) data, even a digraph on it modeling the complex network of speech-signal sequences, and apply these topological inputs for smarter learning.

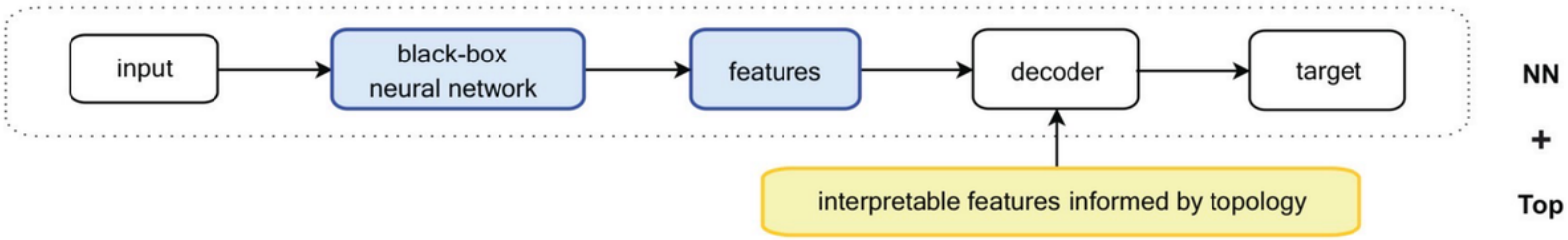
- In a related direction, based on TopCap, we developed topology-enhanced neural networks.

From topological data analysis to topological deep learning

Motivated by the work of Carlsson and colleagues, we have been investigating analogous questions for speech signals, with the additional tool of time-delay embedding for turning time series data to point clouds in Euclidean spaces, as well as spectrograms as their imagery representations.

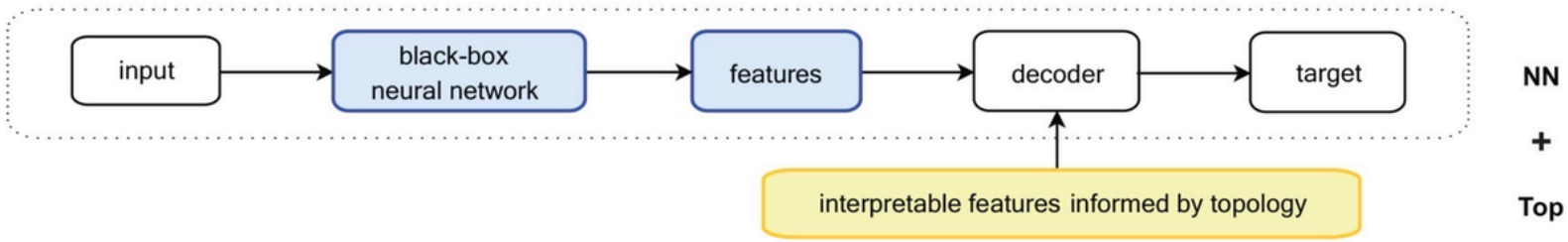
- For phonetic data, linguists created a charted “distribution space” of vowels.
- A main goal remains to use topological methods to reveal a distribution space for speech (and audio) data, even a digraph on it modeling the complex network of speech-signal sequences, and apply these topological inputs for smarter learning.
- In a related direction, based on TopCap, we developed topology-enhanced neural networks.
- Moreover, we exploited the reduced symmetry of spectrograms and designed topological convolutional layers for deep learning speech data.

Topology-enhanced neural networks

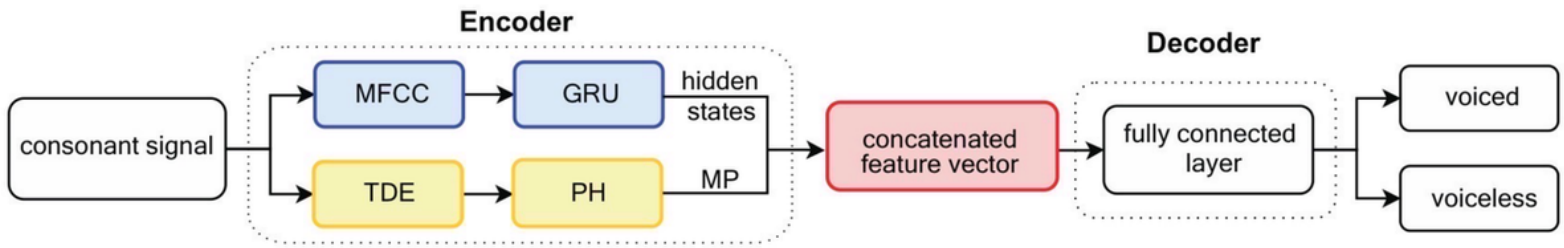


A generic flow chart for enhancing neural networks with topological features

Topology-enhanced neural networks

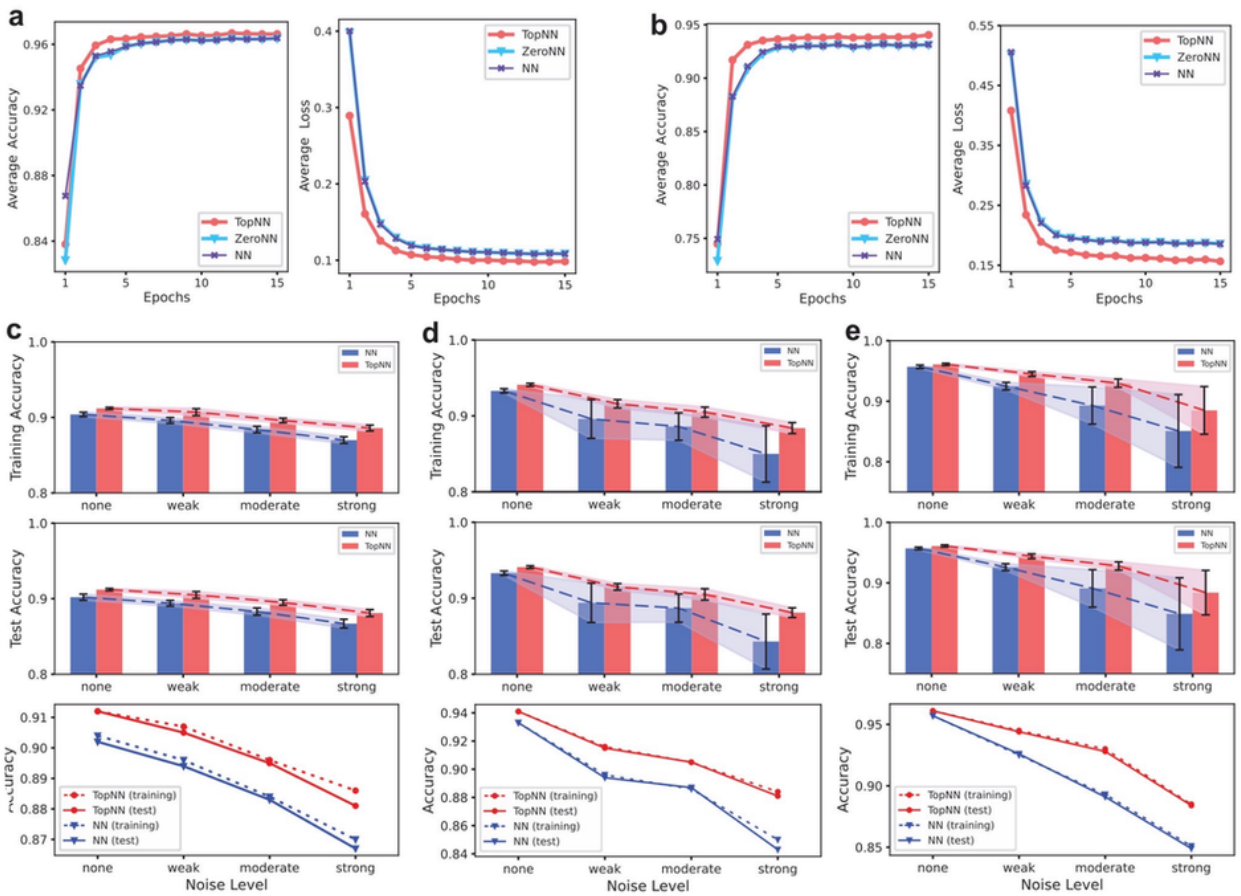


A generic flow chart for enhancing neural networks with topological features



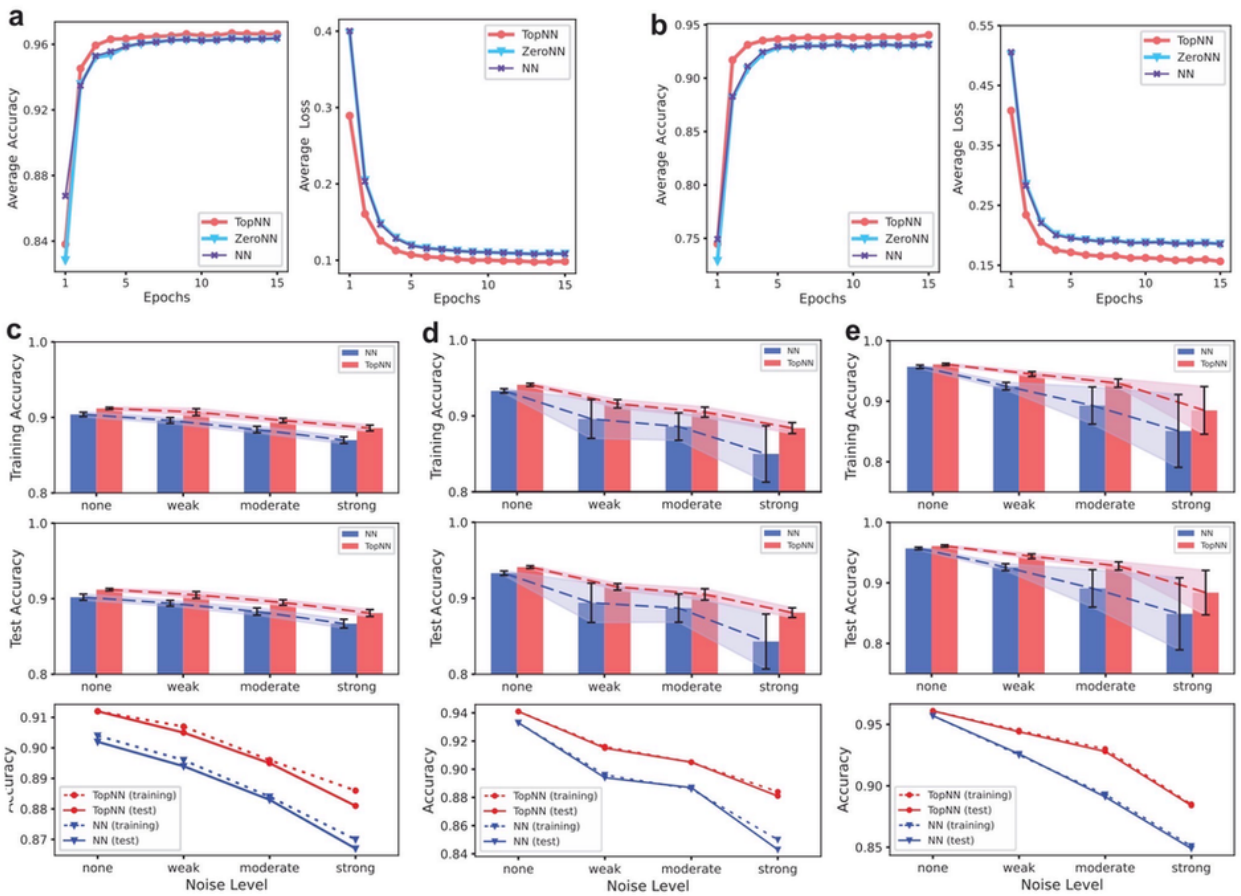
Architecture of a specific TopNN, concatenating GRU and TopCap features

Topology-enhanced neural networks



Visual analytics of experiments with TopNN

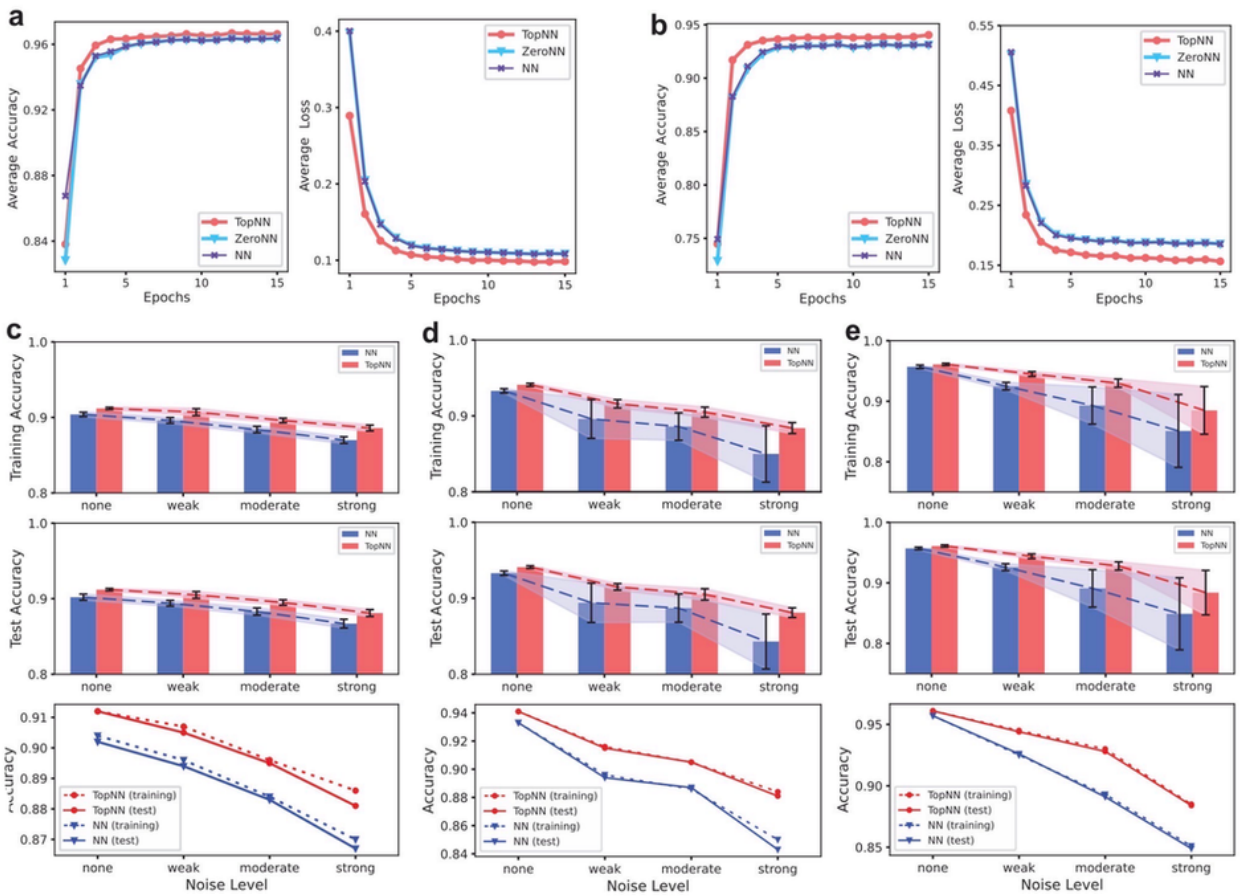
Topology-enhanced neural networks



Visual analytics of experiments with TopNN

***a**, Training curves of TopNN, ZeroNN (NN features concatenated with null topological feature, as a sanity check), and NN on 36000 original speech data from the TIMIT dataset.*

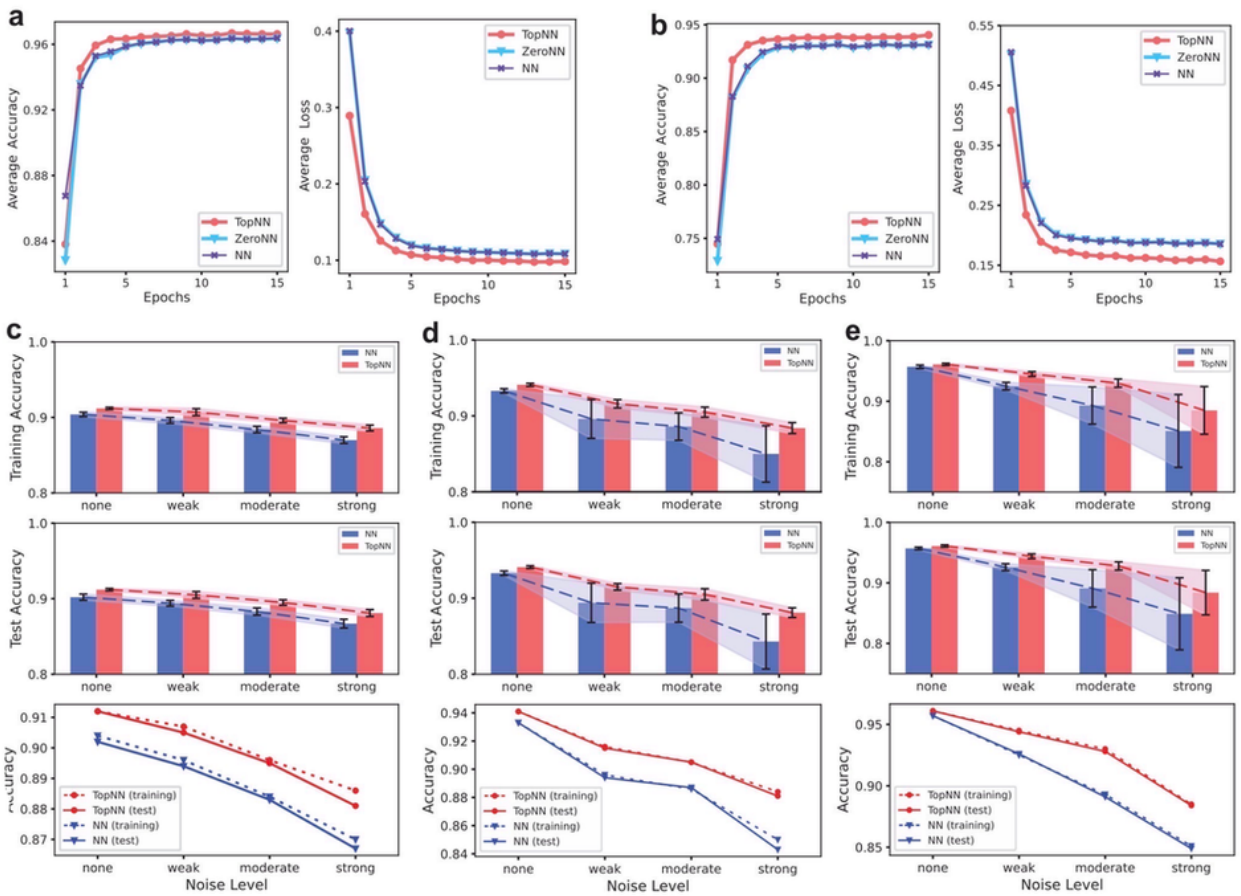
Topology-enhanced neural networks



Visual analytics of experiments with TopNN

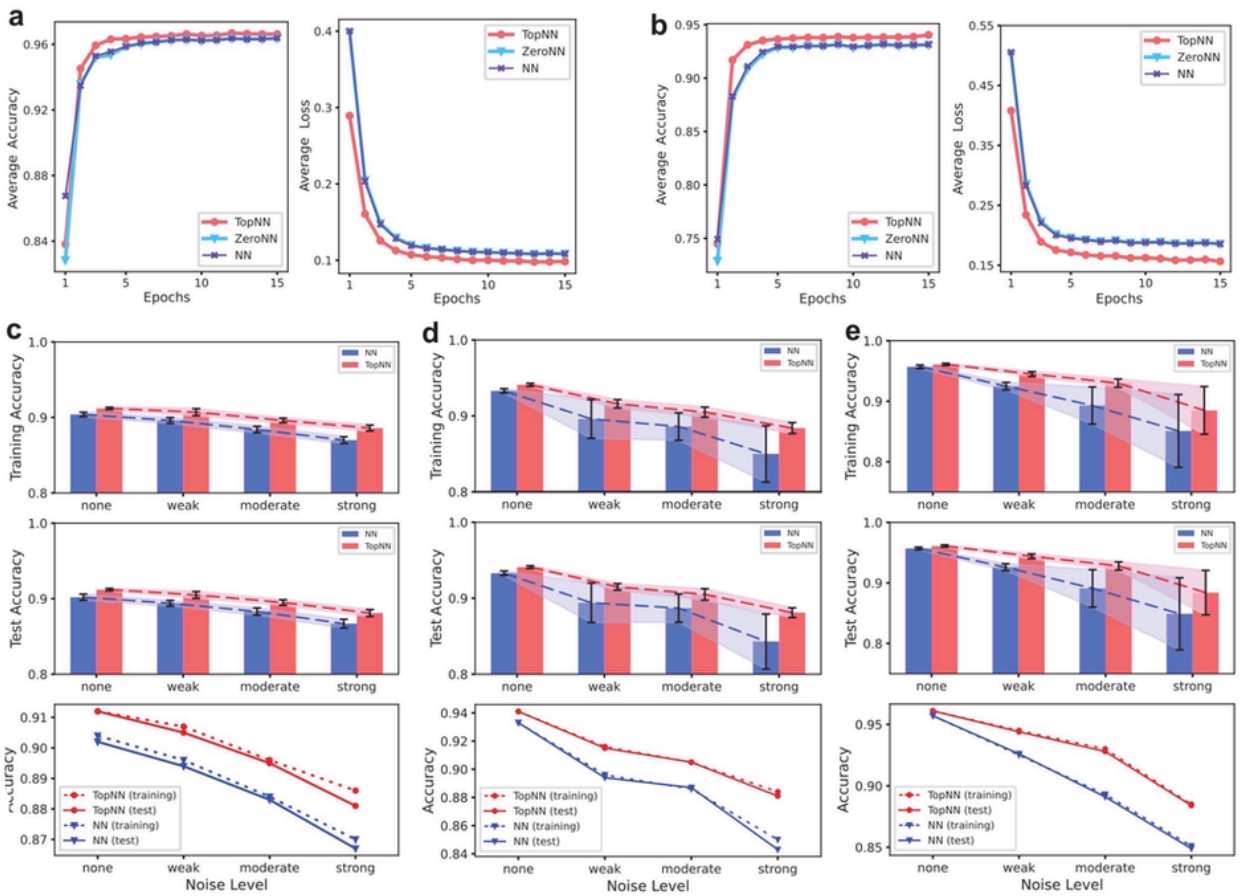
***a**, Training curves of TopNN, ZeroNN (NN features concatenated with null topological feature, as a sanity check), and NN on 36000 original speech data from the TIMIT dataset. They demonstrate that TopNN has higher accuracy and faster convergence in loss function than ZeroNN and NN.*

Topology-enhanced neural networks



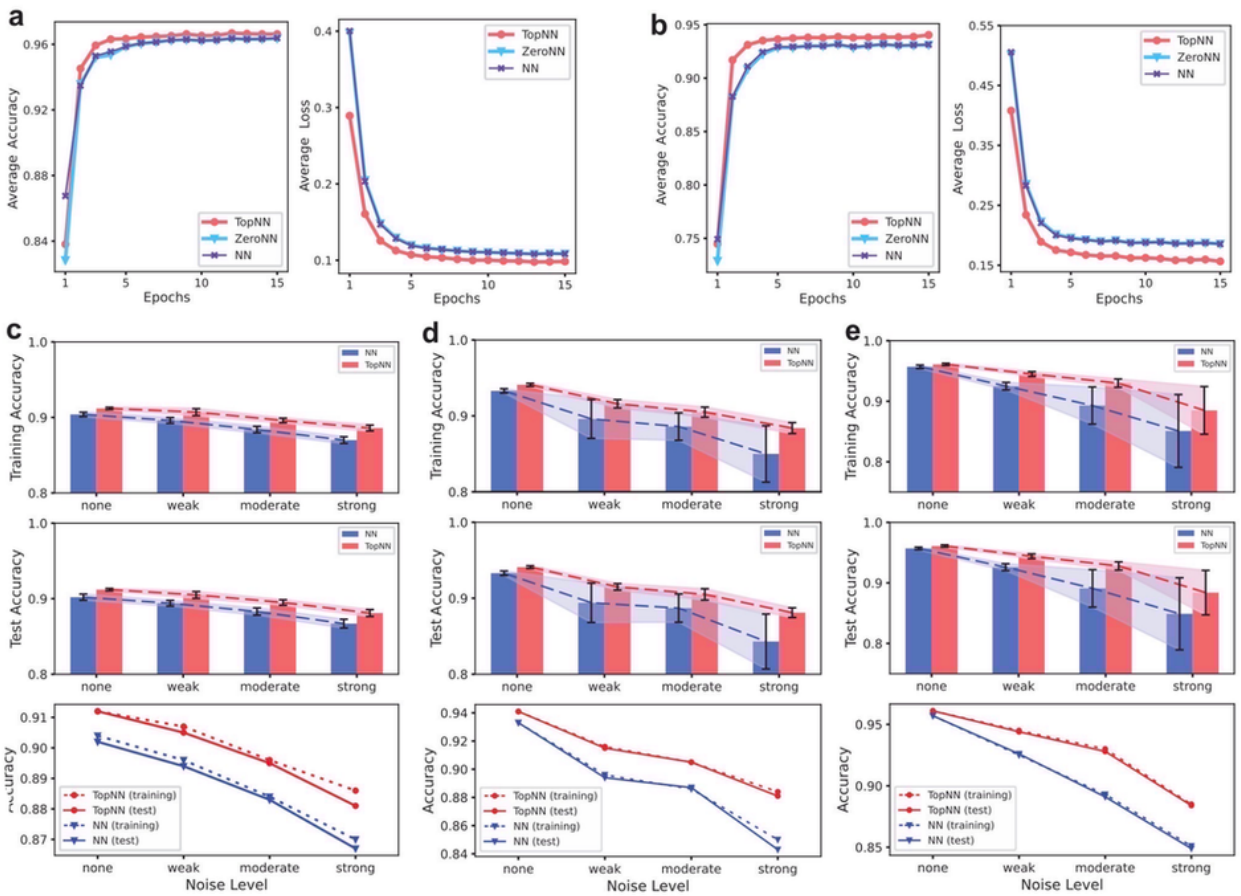
Visual analytics of experiments with TopNN
***b**, Training curves of TopNN, ZeroNN, and NN with the same set up as in **a** and including noise (signal-to-noise ratio = 5dB).*

Topology-enhanced neural networks



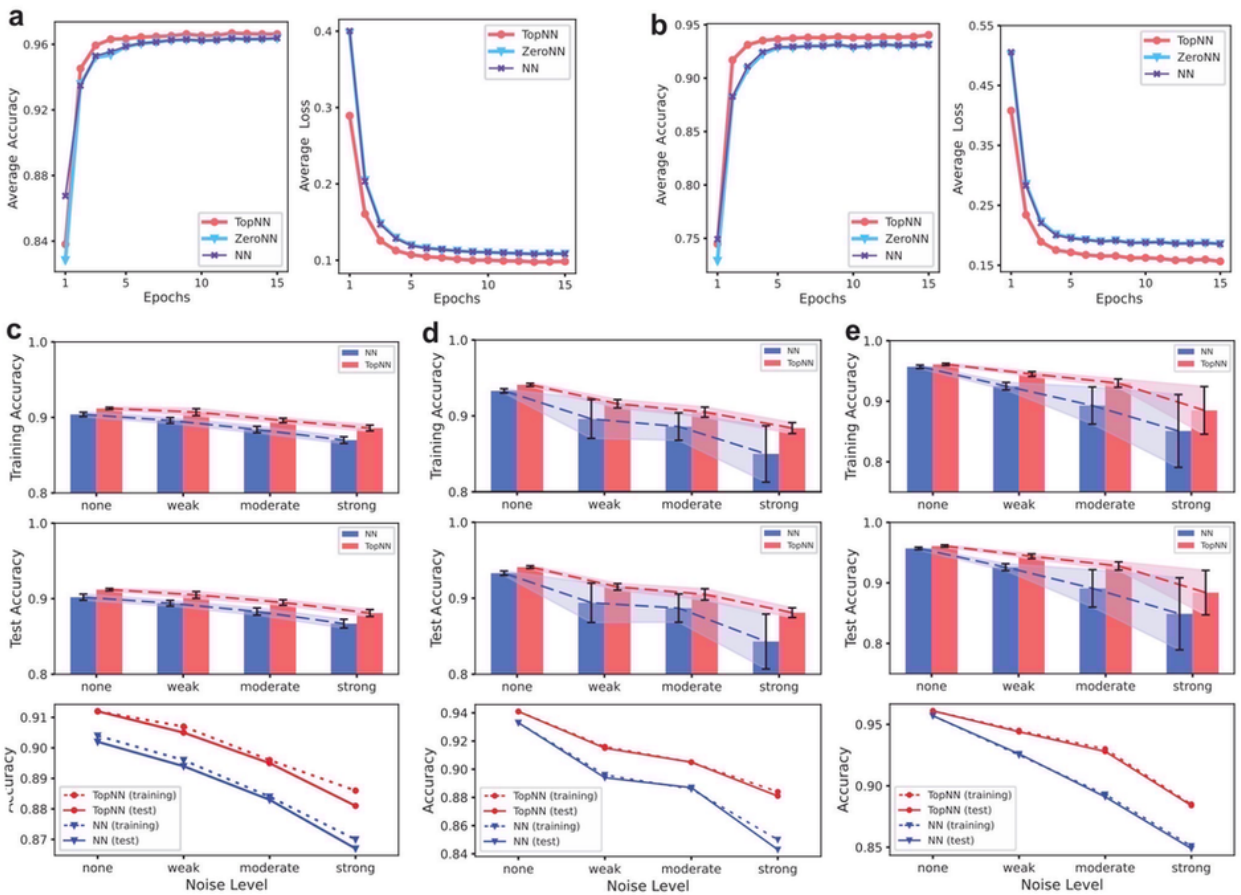
Visual analytics of experiments with TopNN
***b**, Training curves of TopNN, ZeroNN, and NN with the same set up as in **a** and including noise (signal-to-noise ratio = 5dB). With Gaussian white noise added, TopNN's improvement in accuracy and loss decrease are more prominent compared with the results in **a**.*

Topology-enhanced neural networks



Visual analytics of experiments with TopNN
c, ***d***, *and* ***e***, *Comprehensive performance comparison and noise robustness analysis of TopNN and NN based on training and test accuracy rates with the large datasets ALLSSTAR, LJSpeech, and TIMIT, respectively.*

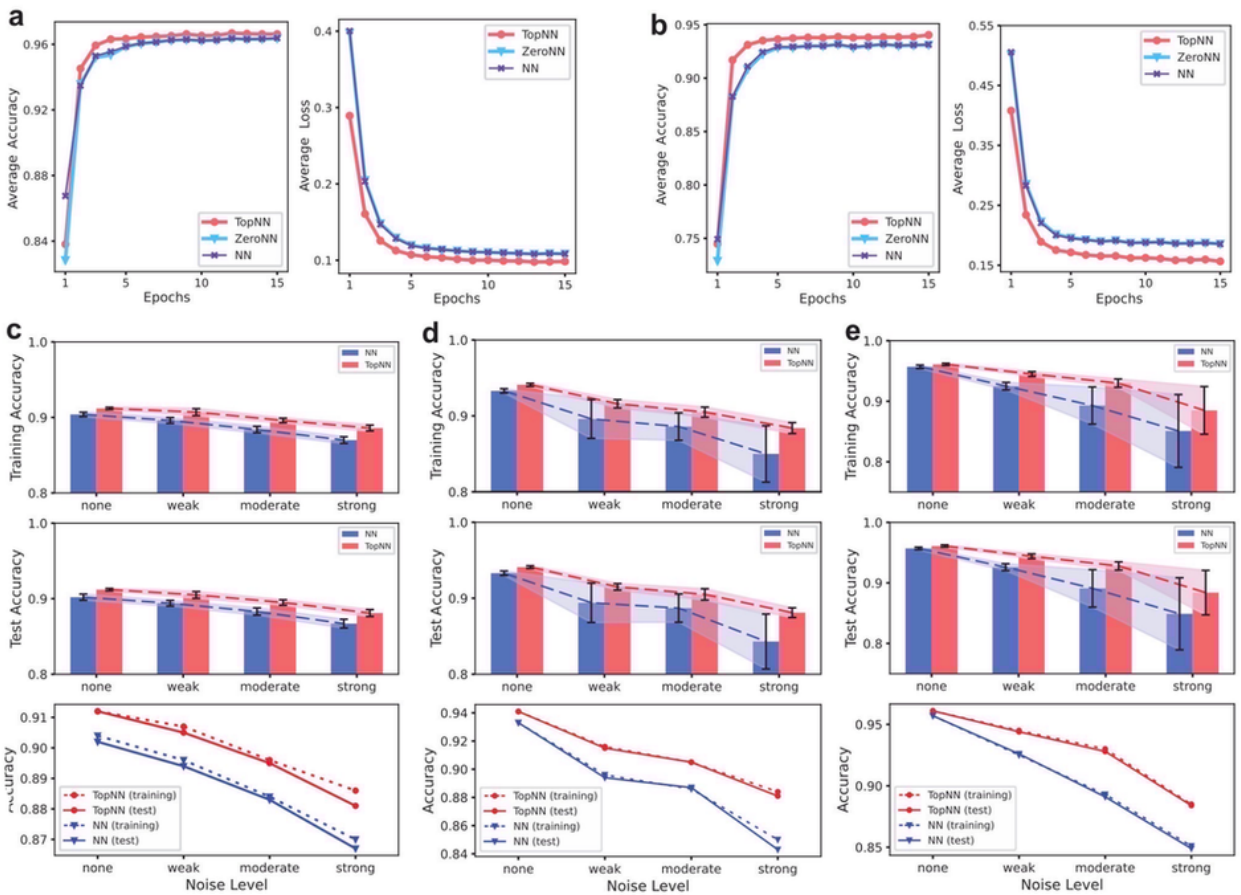
Topology-enhanced neural networks



Visual analytics of experiments with TopNN

c, ***d***, *and* ***e***, *Comprehensive performance comparison and noise robustness analysis of TopNN and NN based on training and test accuracy rates with the large datasets ALLSSTAR, LJSpeech, and TIMIT, respectively. Noise levels include none, weak (SNR = 10dB), moderate (SNR = 5dB), and strong (SNR = 0dB).*

Topology-enhanced neural networks

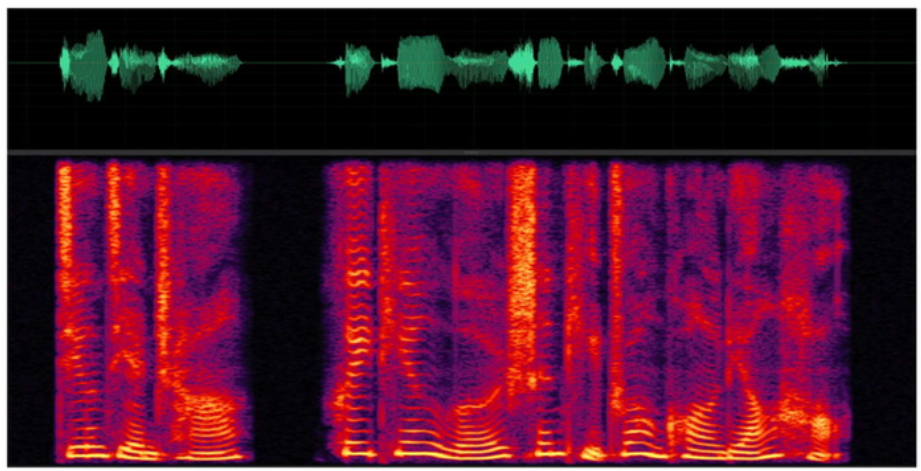


Visual analytics of experiments with TopNN

c, ***d***, *and* ***e***, *Comprehensive performance comparison and noise robustness analysis of TopNN and NN based on training and test accuracy rates with the large datasets ALLSSTAR, LJSpeech, and TIMIT, respectively. Noise levels include none, weak (SNR = 10dB), moderate (SNR = 5dB), and strong (SNR = 0dB). TopNN consistently achieves higher accuracy, steadier performance, and more robustness against noise.*

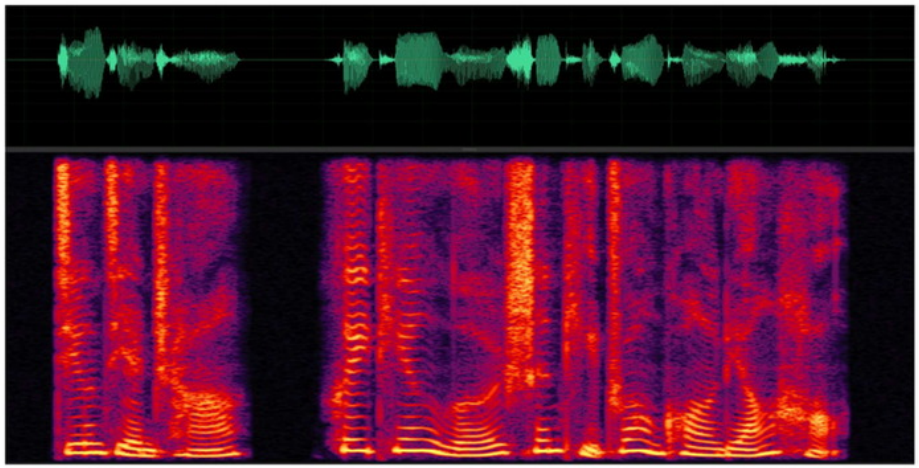
Topology-informed convolution kernels for speech recognition

We defined a notion of *contrast* for 3×3 convolution kernels that process spectrograms, and introduced rigid constraints (unit norm and zero-sum of column vectors) to define a space V of kernels.



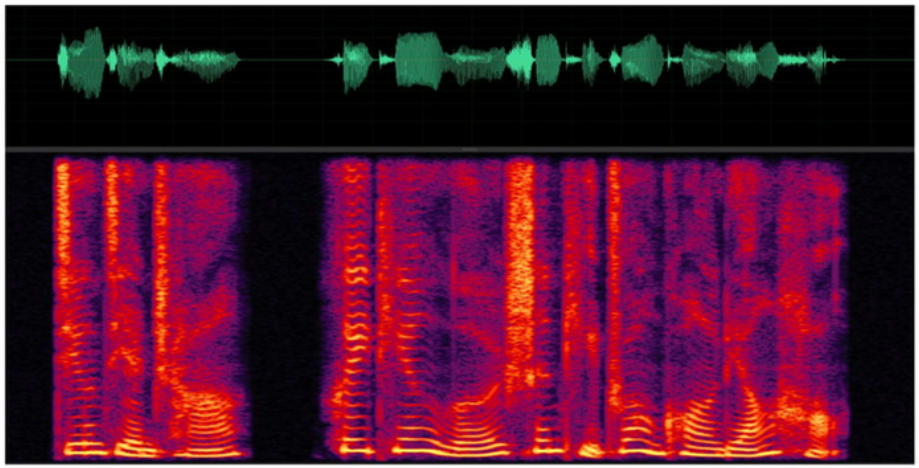
Topology-informed convolution kernels for speech recognition

We defined a notion of *contrast* for 3×3 convolution kernels that process spectrograms, and introduced rigid constraints (unit norm and zero-sum of column vectors) to define a space V of kernels. We showed that V is homeomorphic to S^5 and that the natural $\mathrm{SO}(3)$-action on V induces a quotient space B that is homeomorphic to a disk D^2.

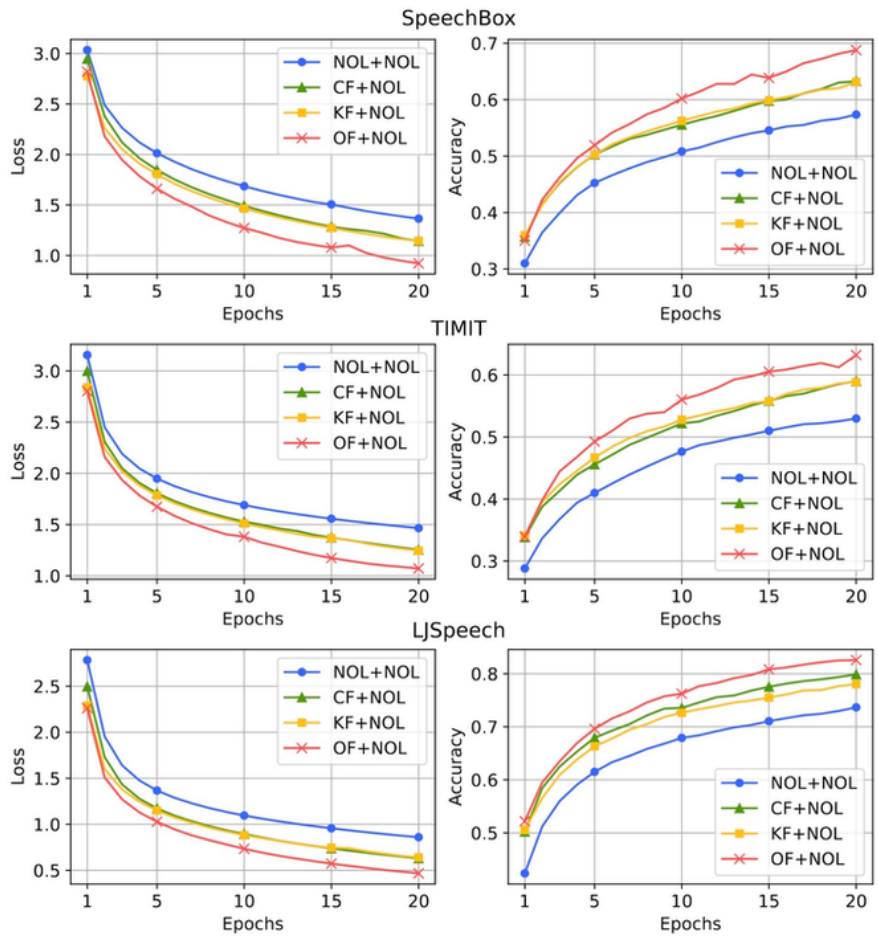


Topology-informed convolution kernels for speech recognition

We defined a notion of *contrast* for 3×3 convolution kernels that process spectrograms, and introduced rigid constraints (unit norm and zero-sum of column vectors) to define a space V of kernels. We showed that V is homeomorphic to S^5 and that the natural $\mathrm{SO}(3)$-action on V induces a quotient space B that is homeomorphic to a disk D^2. We then defined untrained Orthogonal Filter (OF) layer with convolution kernels informed by this topology.

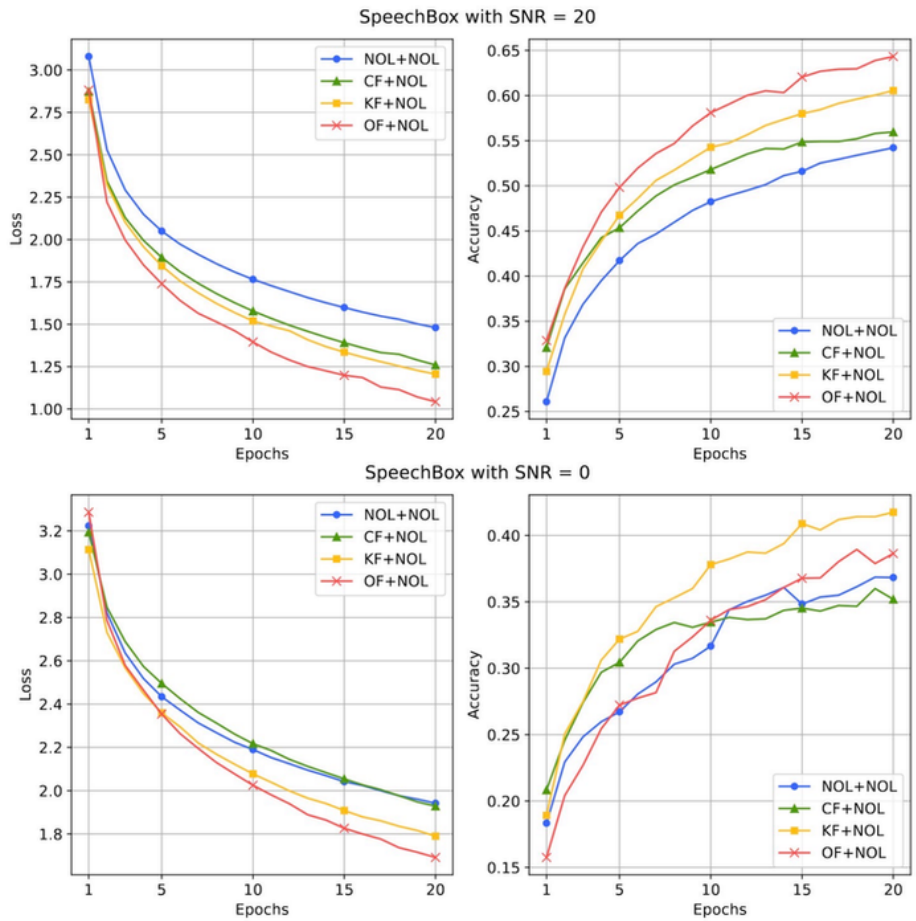


Topology-informed convolution kernels for speech recognition



Comparisons of normal (NOL), Love, Filippenko, Maroulas & Carlsson's circle filter (CF) and Klein-bottle filter (KF), and our orthogonal filter (OF) convolutional layers for phoneme classification tasks via loss and accuracy on datasets SpeechBox, TIMIT, and LJSpeech

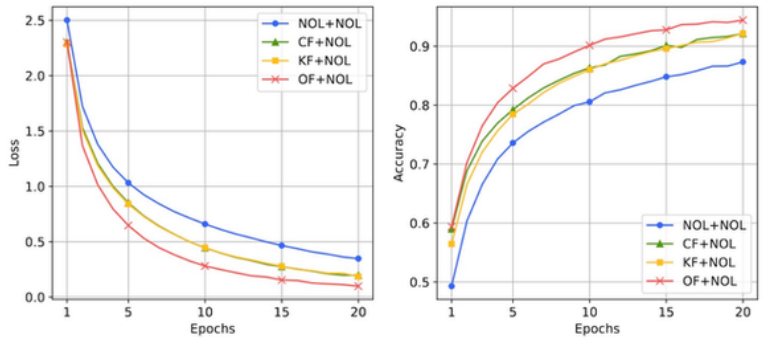
Topology-informed convolution kernels for speech recognition



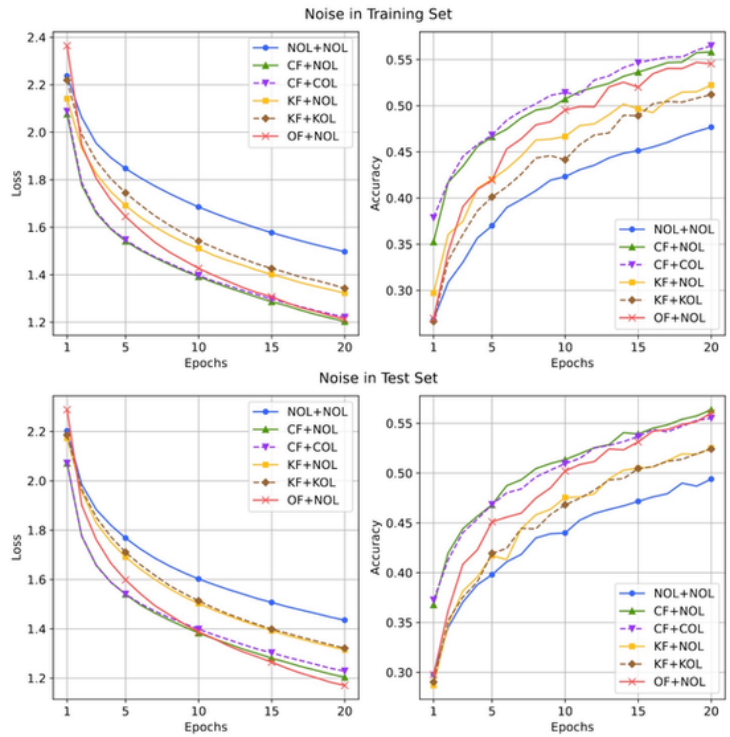
Comparisons with noise added

Our proposed OF layer enables superior performance in phoneme recognition, particularly in low-noise scenarios.

Topology-informed convolution kernels for speech recognition



Comparison for word classification on SpeechCommands

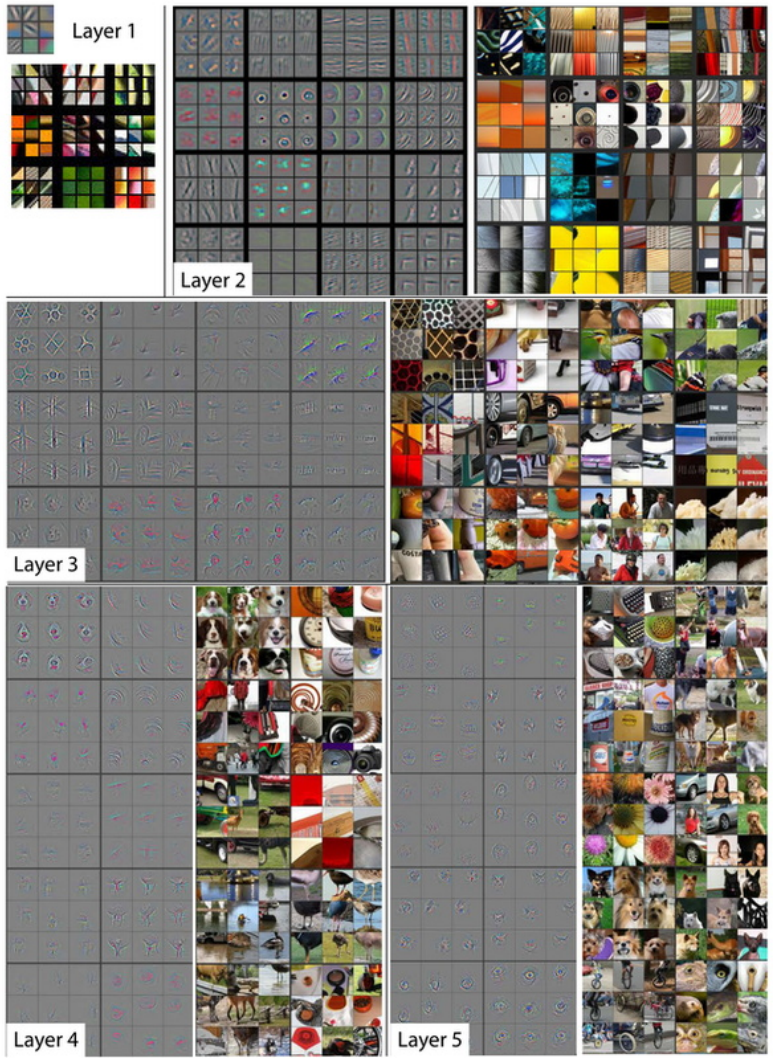


Comparisons for image classification on CIFAR10

Topology-informed convolution kernels: further discussion

Which convolutional layers are *topological*?

- Inspired by Zeiler and Fergus's work on image networks:
  - Shallow layers – local edges, textures;
  - Deeper layers – global, class-specific structures.



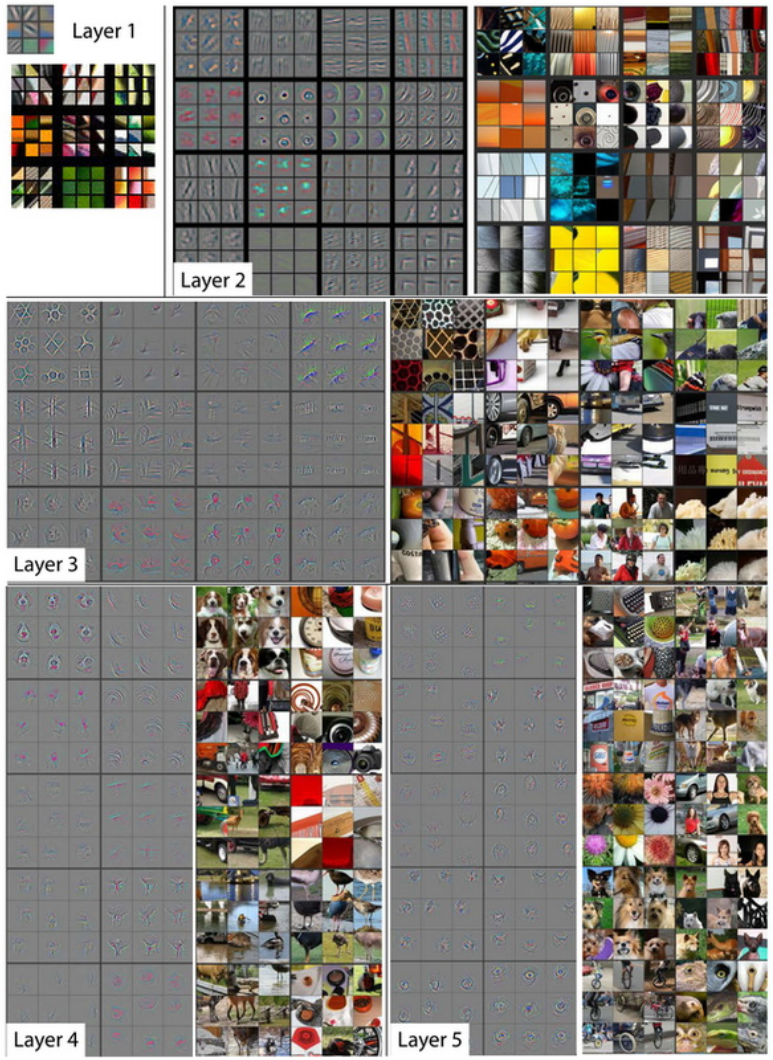
Layer-wise feature visualization

Zeiler & Fergus, *Visualizing and understanding convolutional networks*, **European Conference on Computer Vision**, 2014.

Topology-informed convolution kernels: further discussion

Which convolutional layers are *topological*?

- Inspired by Zeiler and Fergus's work on image networks:
  - Shallow layers – local edges, textures;
  - Deeper layers – global, class-specific structures.
- Love et al.'s Circle and Klein kernels from pixel patches reflect local topology.
- Our Orthogonal kernels from spectrograms reflect more global topology.
- Embed them in different layers accordingly?
- Local-to-global recognition / signal processing is also manifested in human neural networks.



Layer-wise feature visualization

Zeiler & Fergus, *Visualizing and understanding convolutional networks*, **European Conference on Computer Vision**, 2014.

Outlook

1. **[Data Science]** Topological structure of speech signals:
   - Do speech (and audio) signals possess specific topological properties in terms of distributions and dynamics?
   - How can we effectively detect these structures?

2. **[Artificial Intelligence]** Deep learning of speech signals:
   - During the training and testing of speech signals, do artificial neural networks exhibit specific topological distributions and dynamics?
   - How can we effectively utilize these mechanisms as "topological prior" to improve performance of deep learning? (feature, loss, architecture)

3. **[Neuroscience]** Auditory perception of speech signals:
   - In response to speech signals (spontaneous & evoked conditions), do the distribution and dynamic properties of neuronal firing in the auditory cortex of the brain exhibit specific topological structures?
   - How can we understand these mechanisms?

Thank you.

SUSTech Applied & Computational Topology Seminar

https://sustech-topology.github.io/acts