Persistent Homology
An Introduction and a New Text Representation for Natural Language Processing

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

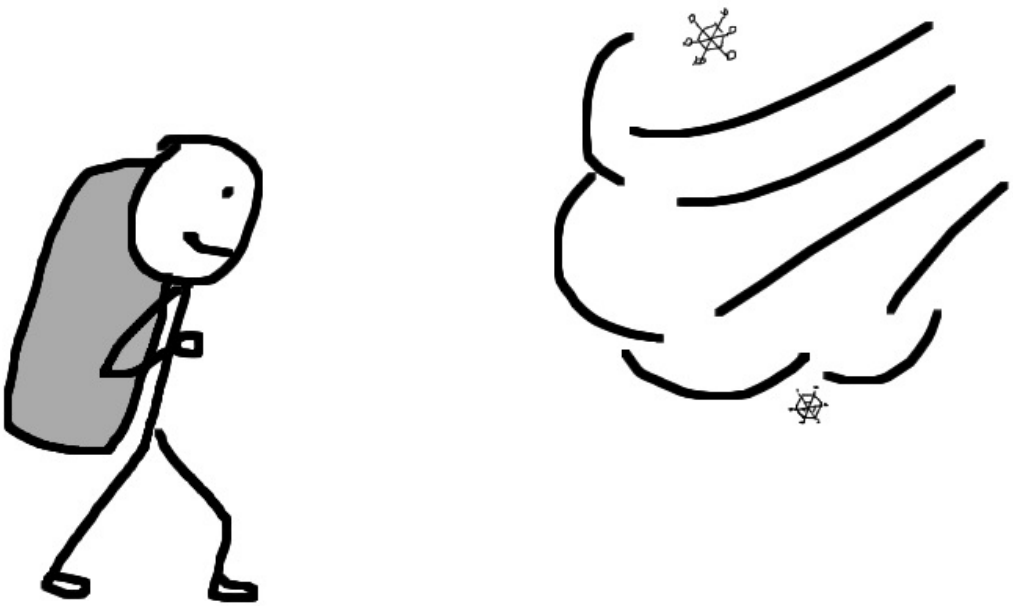
IJCAI 2013

A True Story

Once upon a time, there was a professor in Pittsburgh, who drove 500 miles to the Smoky mountains and started hiking.

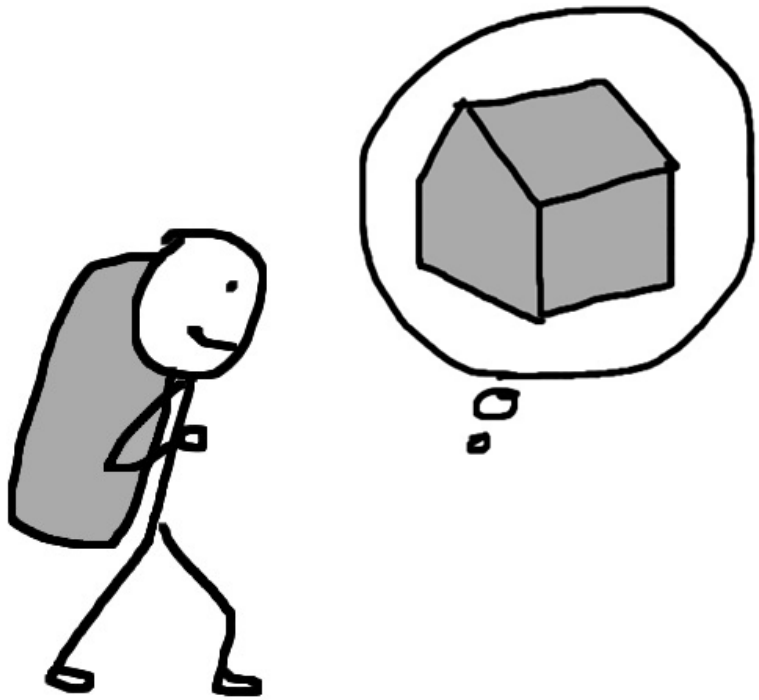
A True Story

It was winter, and the winds were fiercer than he thought. "No problem," he said to himself, for he had reserved a cabin at the end of the trail by phone.



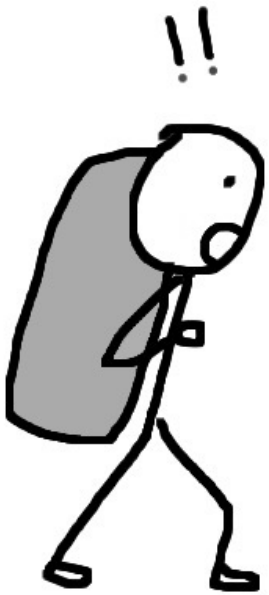
A True Story

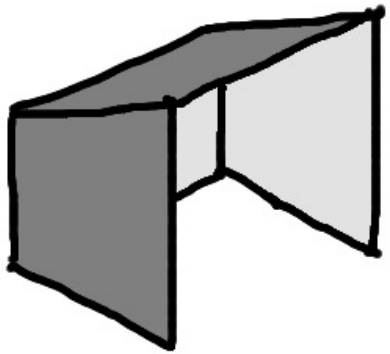
The hike was long. He was getting very cold as the day went by, but was warmed at the thought of the cabin waiting for him.



A True Story

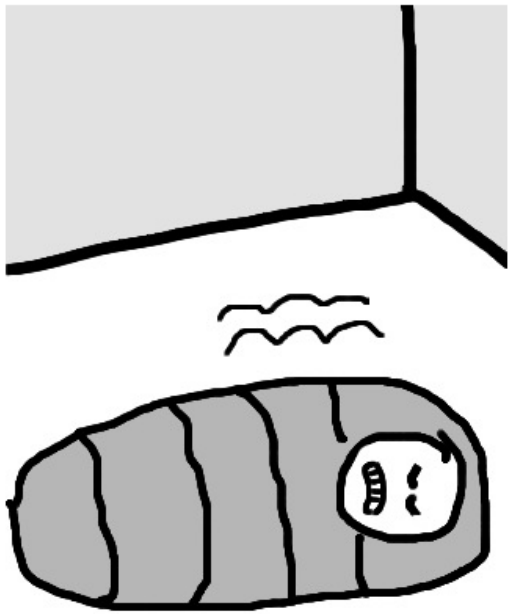
At last he came to the cabin and found out that it had only three walls, instead of four!





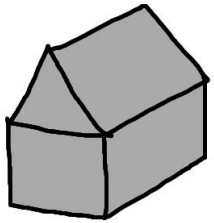
A True Story

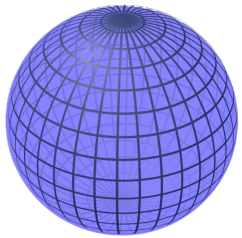
He had to sleep with down jacket inside his thermal sleeping bag.

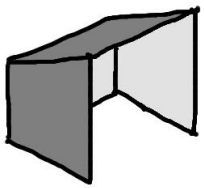


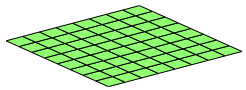
A True Story

The Moral of the story: homology is important.

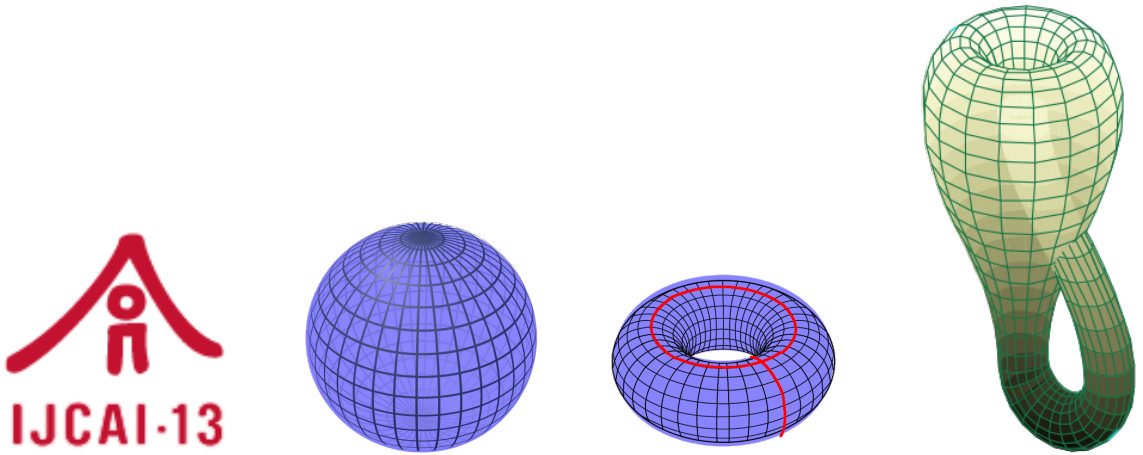


homeomorphic to  , $\text{Betti}_2 = 1$



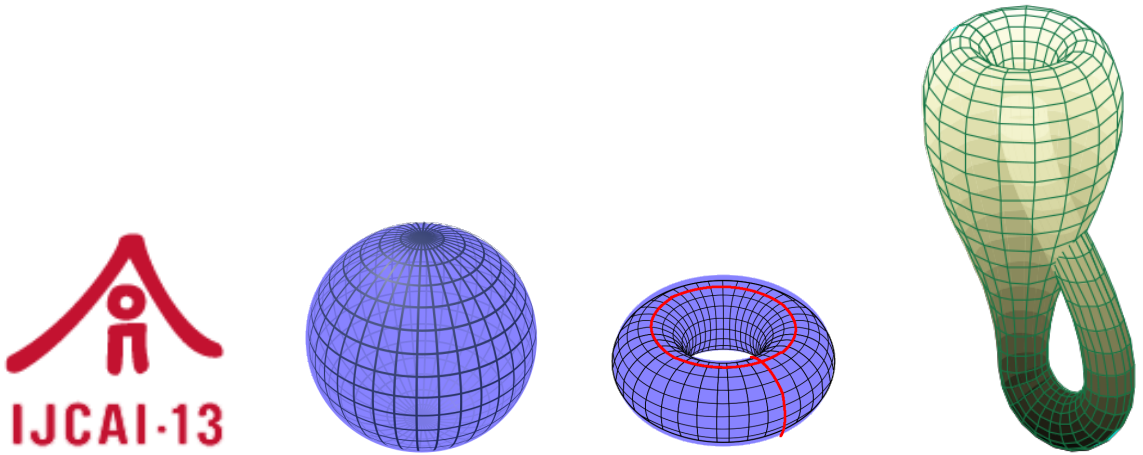
homeomorphic to  , $\text{Betti}_2 = 0$

Homology



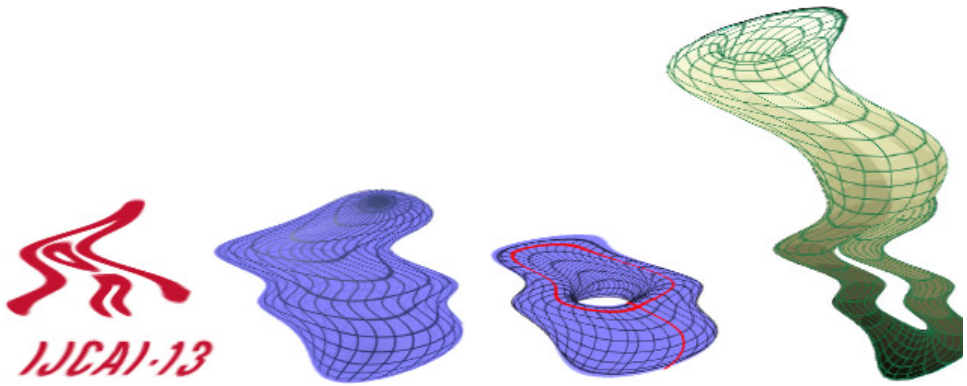
Betti_0 (clusters)	11	1	1	1
Betti_1 (holes)	2	0	2	2
Betti_2 (voids)	0	1	1	1

Homology



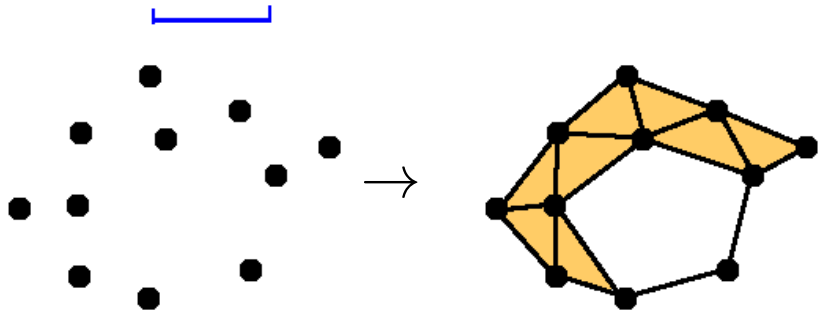
Betti_0 (clusters)	11	1	1	1
Betti_1 (holes)	2	0	2	2
Betti_2 (voids)	0	1	1	1

Homology as features: insensitive to certain distortions

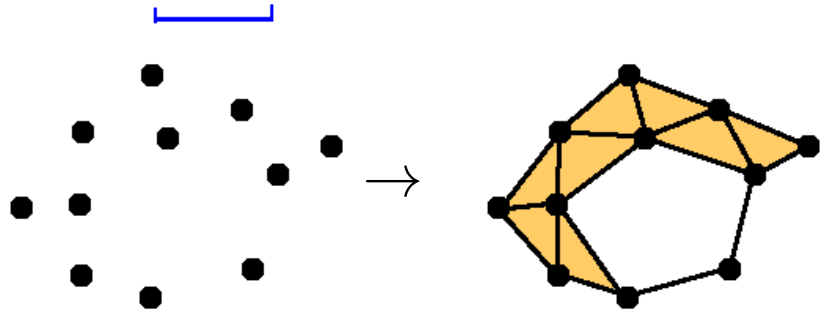


From data to homology

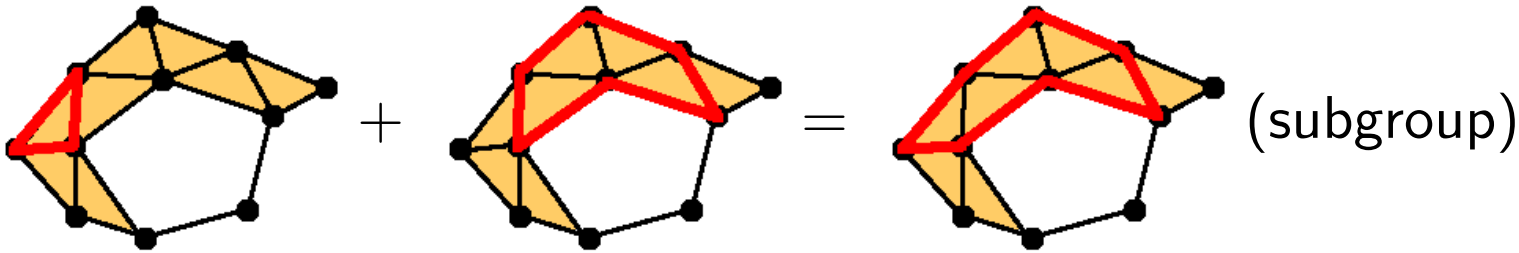
Vietoris-Rips complex

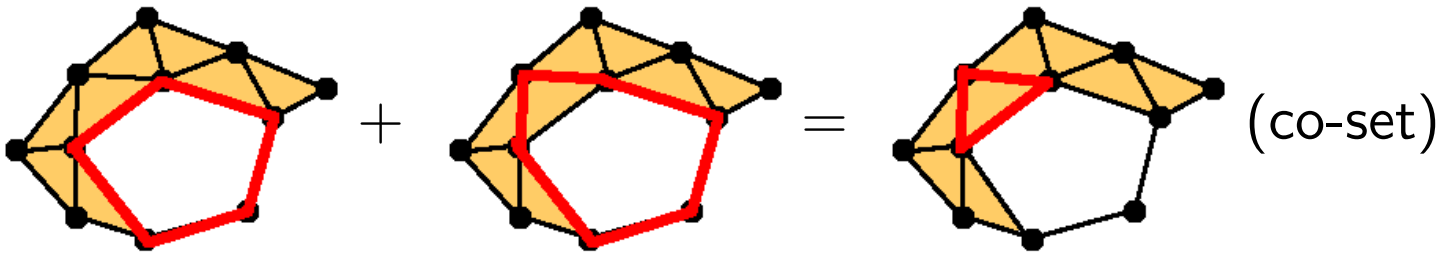


From data to homology

Vietoris-Rips complex 

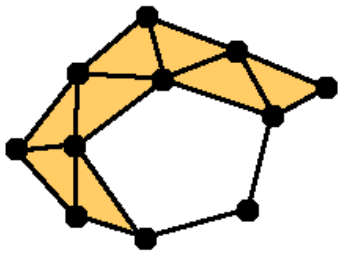
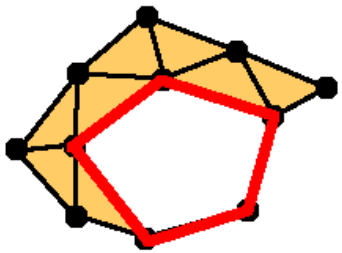
Cycle group (mod 2 addition)





From data to homology

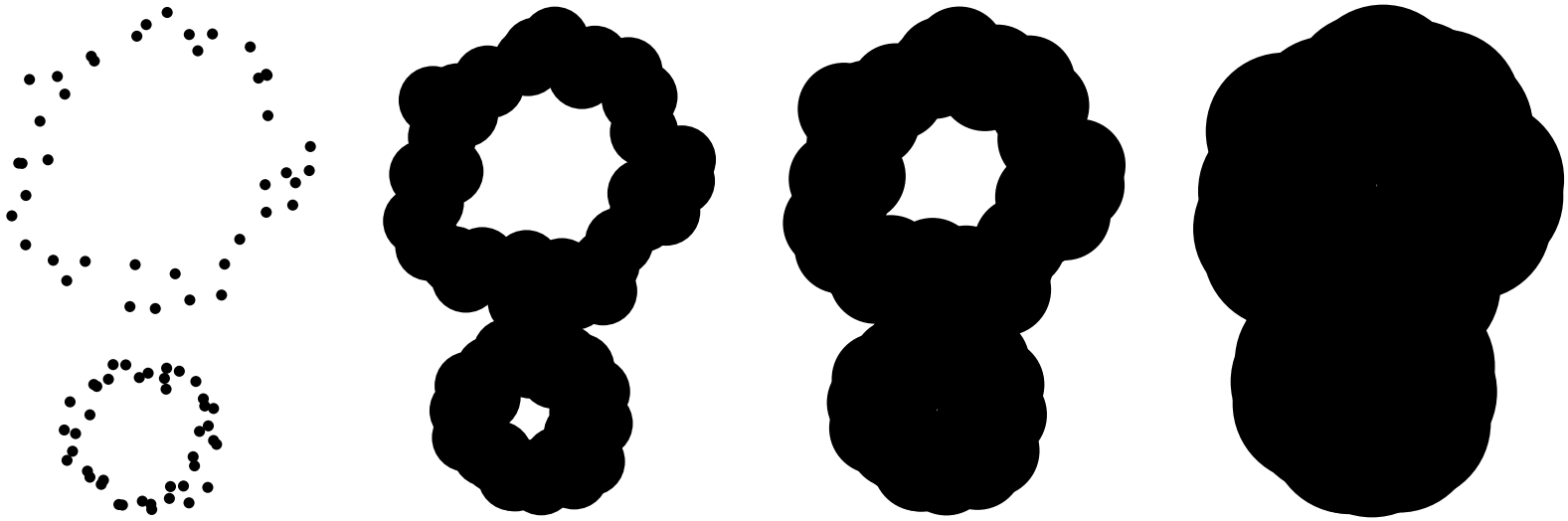
$$\text{Homology } H_1 = \text{quotient group of } \frac{\text{cycle group}}{\text{boundary cycle subgroup}} = \frac{\ker \partial_1}{\text{Img} \partial_2}$$

$H_1 = \{$  ,  $\}$

$\text{Betti}_1 = \text{rank}(H_1) = 1$ (one hole)

See paper for technical details!

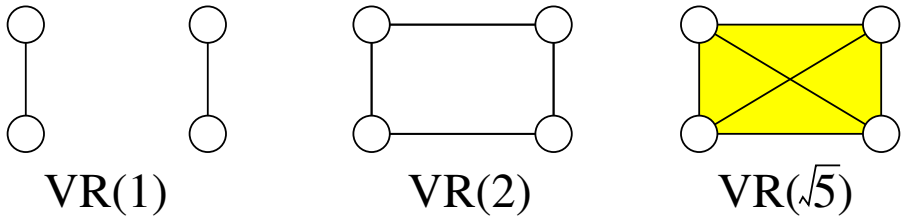
Persistent Homology

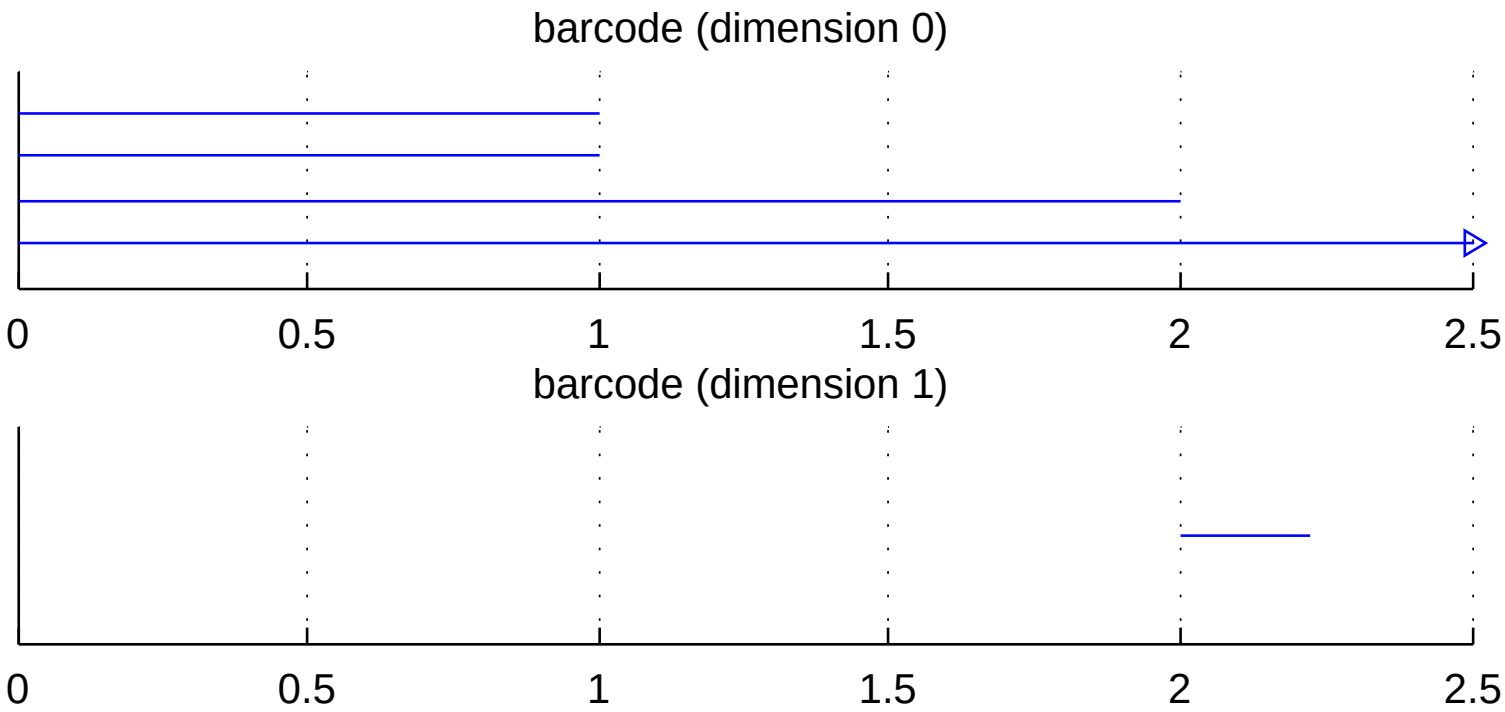


$\text{Betti}_1 =$ 0 2 1 0

Barcode

In persistent homology, when is a hole "born" and when does it "die"?

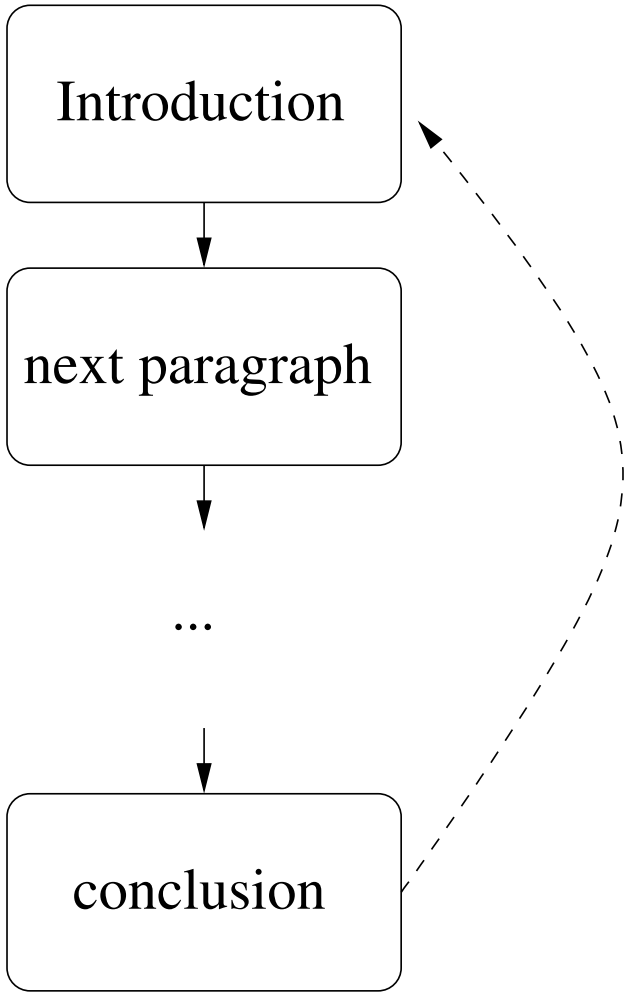




See paper for technical details!

Applications to natural language processing

Some good articles "tie back." Capture such loops with homology.



Example: Itsy bitsy spider

The Itsy Bitsy Spider climbed up the water spout
Down came the rain and washed the spider out
Out came the sun and dried up all the rain

And the Itsy Bitsy Spider climbed up the spout again

- bag-of-words

again	all	and	bitsy	came	climbed	down	dried	itsy	out	rain	spider	spout	sun	the	up	washed	water
0	0	0	1	0	1	0	0	1	0	0	1	1	0	2	1	0	1
0	0	1	0	1	0	1	0	0	1	1	1	0	0	2	0	1	0
0	1	1	0	1	0	0	1	0	1	1	0	0	1	2	1	0	0
1	0	1	1	0	1	0	0	1	0	0	1	1	0	2	1	0	0

- vertices

- tf.idf-based cosine distance

Similarity Filtration (SIF)

$D(x_i, x_j)$ cosine distance between sentences i, j

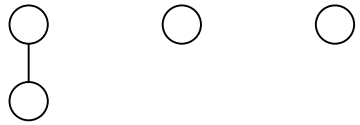
$D_{max} = \max D(x_i, x_j), \forall i, j = 1 \ldots n$
FOR $m = 0, 1, \ldots M$
Add $VR\left(\frac{m}{M} D_{max}\right)$ to the filtration
END
Compute persistent homology on the filtration

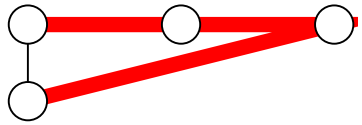
Similarity Filtration with Time Skeleton (SIFTS)

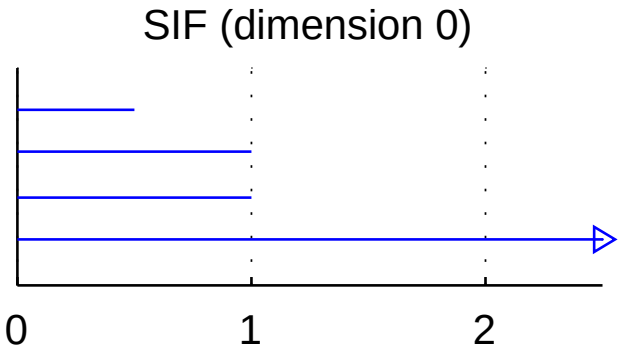
Add time edges

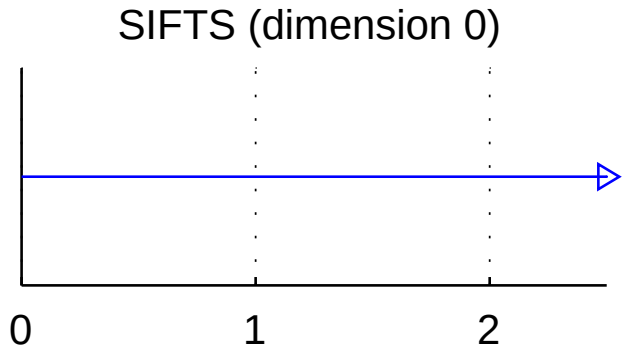
$D(x_i, x_{i+1}) = 0$ for $i = 1, \ldots, n-1$
$D_{max} = \max D(x_i, x_j), \forall i, j = 1 \ldots n$
FOR $m = 0, 1, \ldots M$
Add $VR\left(\frac{m}{M} D_{max}\right)$ to the filtration
END
Compute persistent homology on the filtration

SIF vs. SIFTS on Itsy bitsy spider

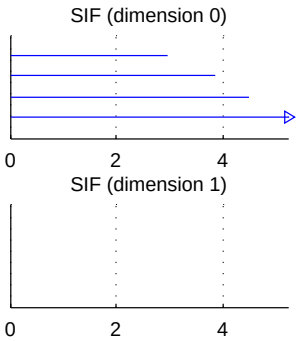


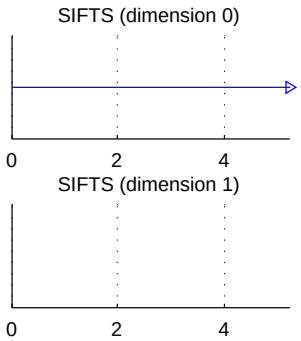




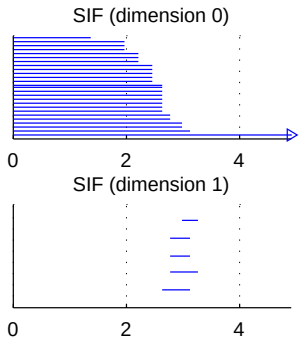


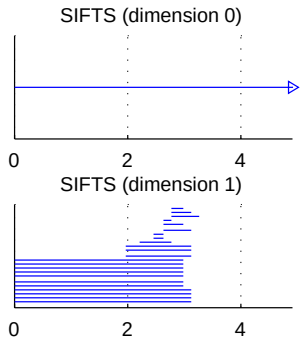
On Nursery Rhymes and Other Stories



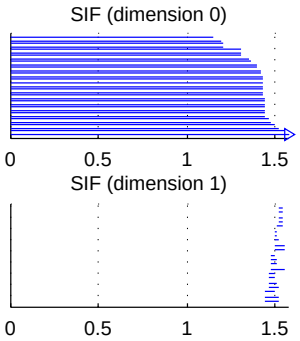


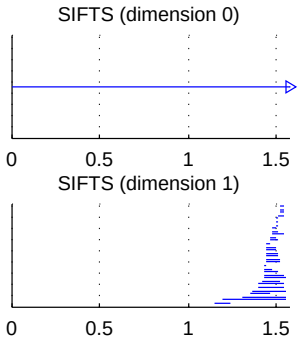
Row Row Row Your Boat



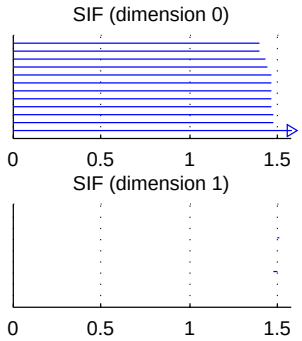


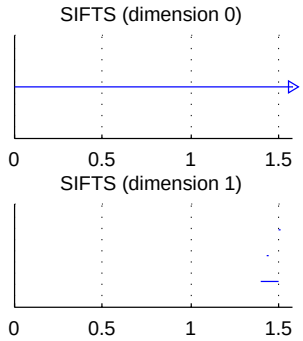
London Bridge





Little Red-Cap





Alice in Wonderland

- London Bridge: "My fair Lady" repeats 12 times.
- Little Red-Cap: "The better to see you with, my dear" and "The better to eat you with!"

On Child and Adolescent Writing

- Older writers have more complex barcodes?
- LUCY corpus: children (ages 9–12, 150 essays), undergraduates (48 essays)
- average article length: child=11.6 sentences, adolescent=25.8
- SIFTS barcode summary statistics:
  - holes?: what percentage of articles have H_1 holes
  - $|H_1|$: number of holes in the article
  - ϵ^*: the smallest ϵ when the first hole in H_1 forms.

	child	adolescent	adol. trunc.
holes?	87%	100%*	98%*
$\lvert H_1\rvert$	3.0 ($\pm$0.2)	17.6 ($\pm$0.9)*	3.9 ($\pm$0.2)*
ϵ^*	1.35 ($\pm$.02)	1.27 ($\pm$.02)*	1.38 ($\pm$.01)

*: statistically significantly different from "child"

Summary

- Persistent homology can offer useful representations for machine learning
- Where is the "killer app"?

Acknowledgments: Pradeep Ravikumar for the story, Kevyn Collins-Thompson for corpus, and US National Science Foundation.