20.4.22

# Topological Insights on Vector-Embedded Language

Noah J. Bergam (Columbia University),
Marian Gidea (Yeshiva University)

REU in Stochastic Processes and Financial Time Series, Yeshiva University, 2021

# Roadmap

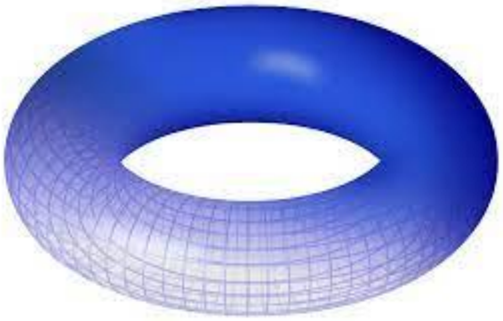
Background

- Topological data analysis
- Word embeddings

Methods

Results + Exploration

Conclusion

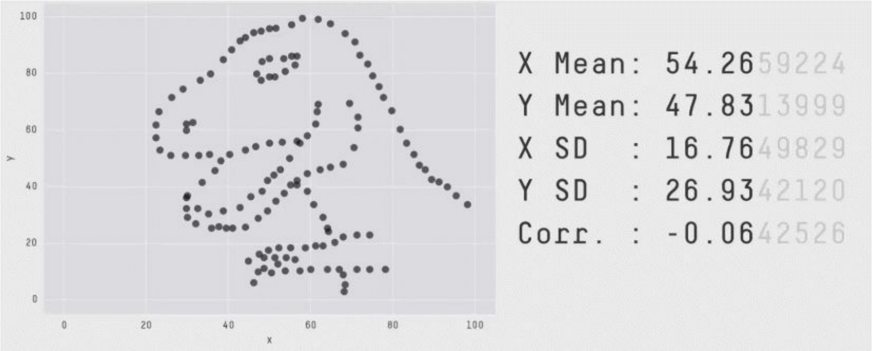


# Topological Data Analysis (TDA)

TDA is a method of **quantifying the shape** of (high-dimensional) data using **persistent homology**

<u>Main idea</u>:

- (1) Consider a ball of radius epsilon around each data point
- (2) Vary epsilon and record when “holes” are “born” or “die”



The limitations of descriptive statistics...

# Pipeline

Point Cloud →

Simplicial Complex →

Homology Group →

**Persistence Diagram →**

Persistence Landscape →

Norm

(In a sense, we go from *statistical* to *topological* back to *statistical*.)

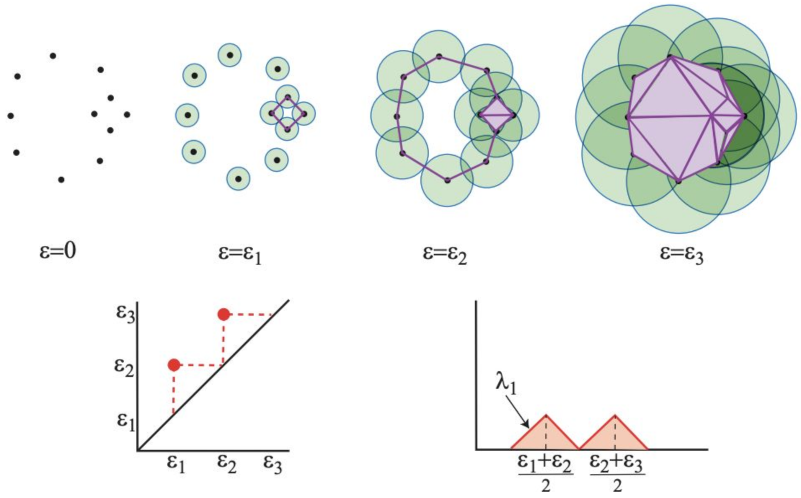


Figure 1: Rips filtration of simplicial complexes illustrating the birth and death of loops; the 1-dimensional persistence diagram and the corresponding persistence landscape are shown below.

$$\|\eta\|_p = \left( \sum_{k=1}^{\infty} \|\eta_k\|_p^p \right)^{1/p}$$

*The L-p norm of the persistence landscape!*

[1]: Gidea M., Katz Y. Topological data analysis of financial time series: Landscapes of crashes. Physica A. 2018;491(1):820–834.

# Word Embeddings

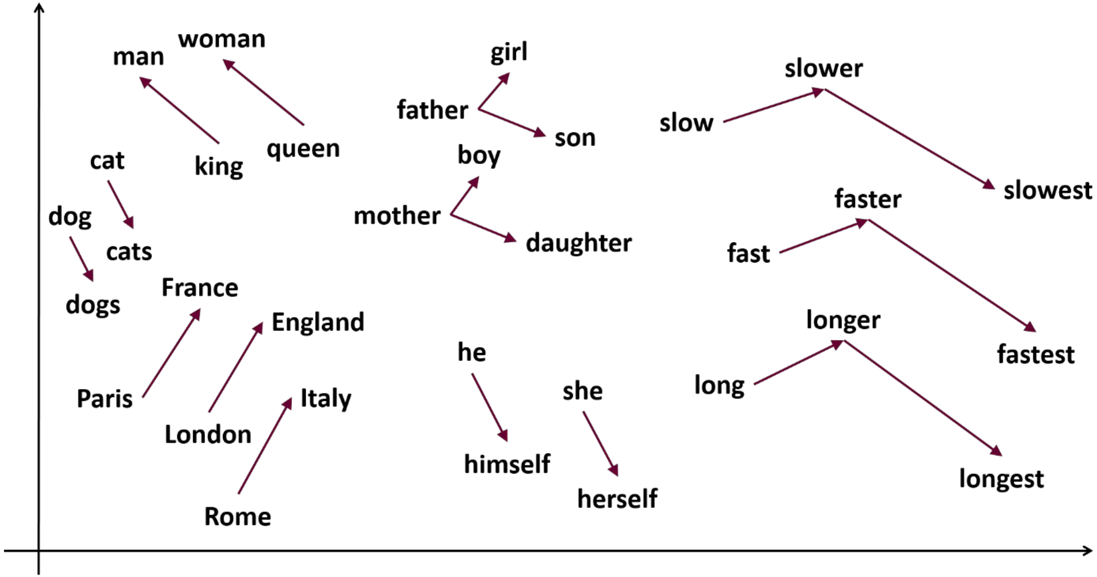
Representing words as dense vectors.

**Classical**: statistical

- SVD

**Modern**: neural

- word2vec (2013)
- GloVE (2014)
- BERT (2018)



Word vectors can capture *analogies*.

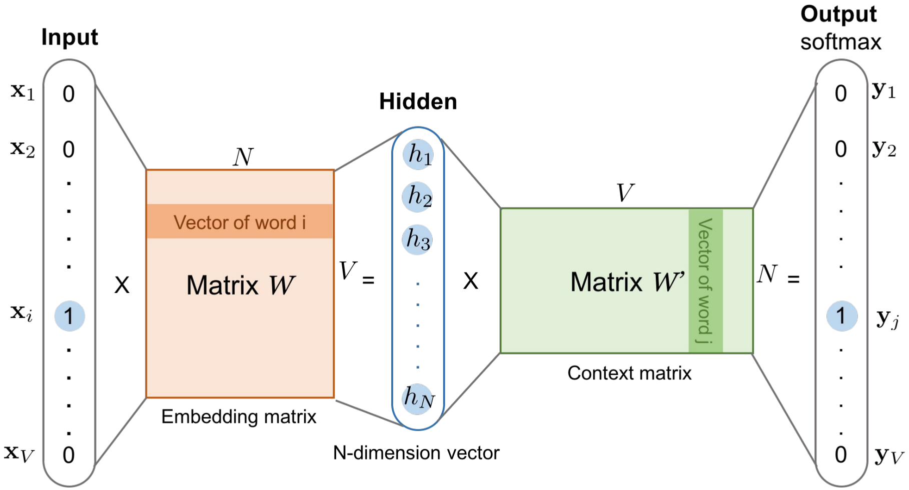
# How word2vec embeddings work

Keep track of a word’s “context”

- e.g. i ate broccoli and i somehow enjoyed it thoroughly

Train a neural network to predict **P(context | word)** over all words and contexts.

- Elegantly self-supervised.
- Use W for the word embeddings



# Motivation

Word embeddings allow us to think about text spatially (and thus topologically)

- Language is a “trajectory” in vocabulary space
- Analyzing that trajectory can yield insights on style, intent, plot structure, etc.
- TDA can help us…

**Hypotheses**:

- Higher Lp norm suggests generally larger diversity of vocabulary and therefore some sense of creativity
- Monitoring the Lp norm can help us distinguish the styles between authors.

# Experiment

Consider the “Spooky Author Identification” dataset (Kaggle)

- **Task**: Predict whether a given excerpt was written by Edgar Allan Poe, HP Lovecraft, or Mary Shelley.

**Our Method:** Convert text into a sequence of word embeddings. Split sequence into n “point clouds.” Apply LP norm to each cloud. Use this as ML feature.

**Compare** with max, min, and sum pooling of word embeddings.

We limit ourselves to **Logistic Regression** (simplest classifier)

# Results

Looked at the number of clouds (n) for TDA implementation

- Otherwise, no general hyperparameter search

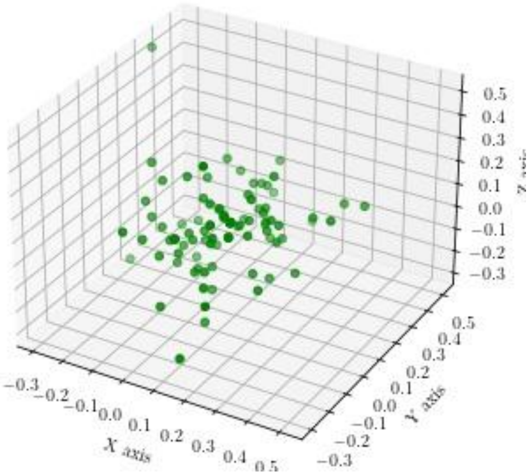
Note that bag-of-words had much higher accuracy (maybe word vectors were not best for this task…)

**Next slides**: Some other exploration…

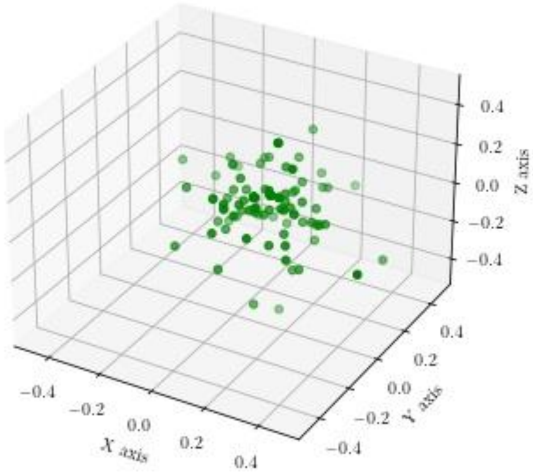
| | Accuracy |
|---|---|
| Bag-of-words | **0.729** |
| Word Vectors MAX pool | 0.489 |
| Word Vectors MIN pool | 0.406 |
| Word Vectors SUM pool | 0.508 |
| Word Vectors TDA (n = 5) | 0.428 |
| Word Vectors TDA (n = 30) | 0.442 |
| Word Vectors TDA (n = 50) | 0.486 |
| Word Vectors TDA (n = 70) | **0.522** |

# Exploration / Visualization

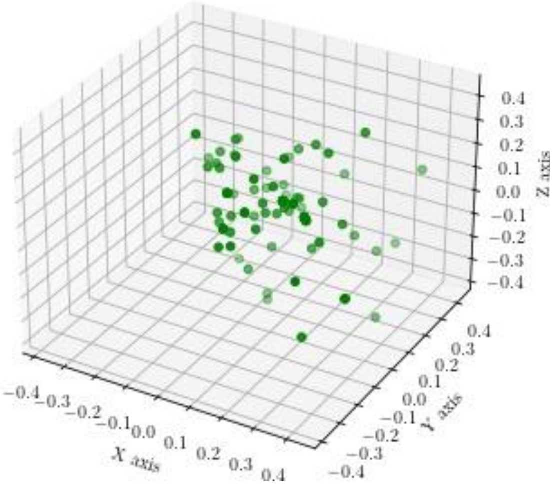
Howl, by Allen Ginsberg

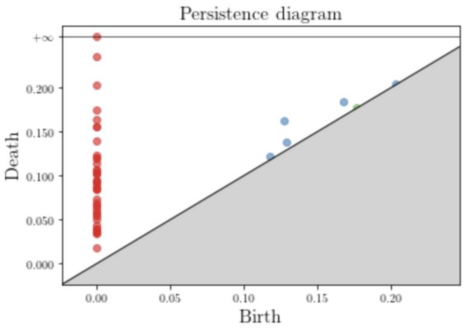


The Raven, by Edgar Allan Poe

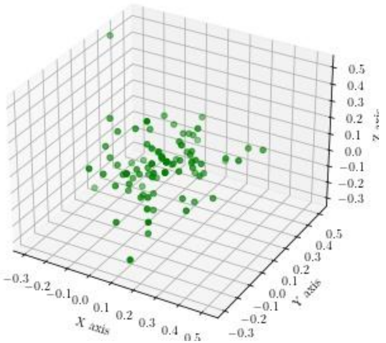


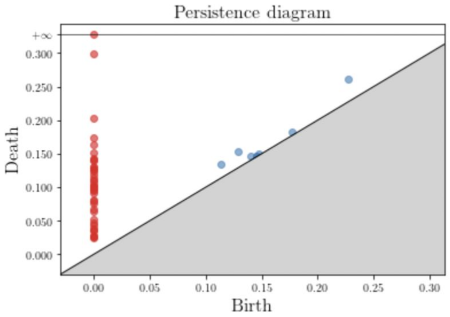
O Captain, My Captain,by Walt Whitman



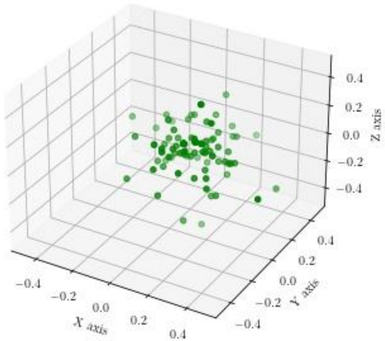


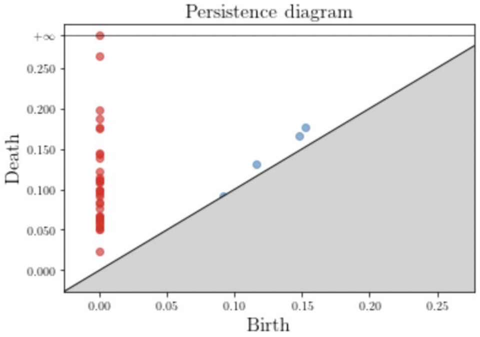
Howl, by Allen Ginsberg



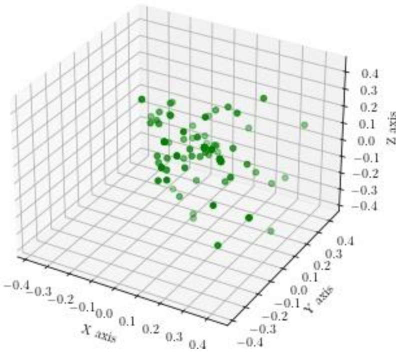


The Raven, by Edgar Allan Poe

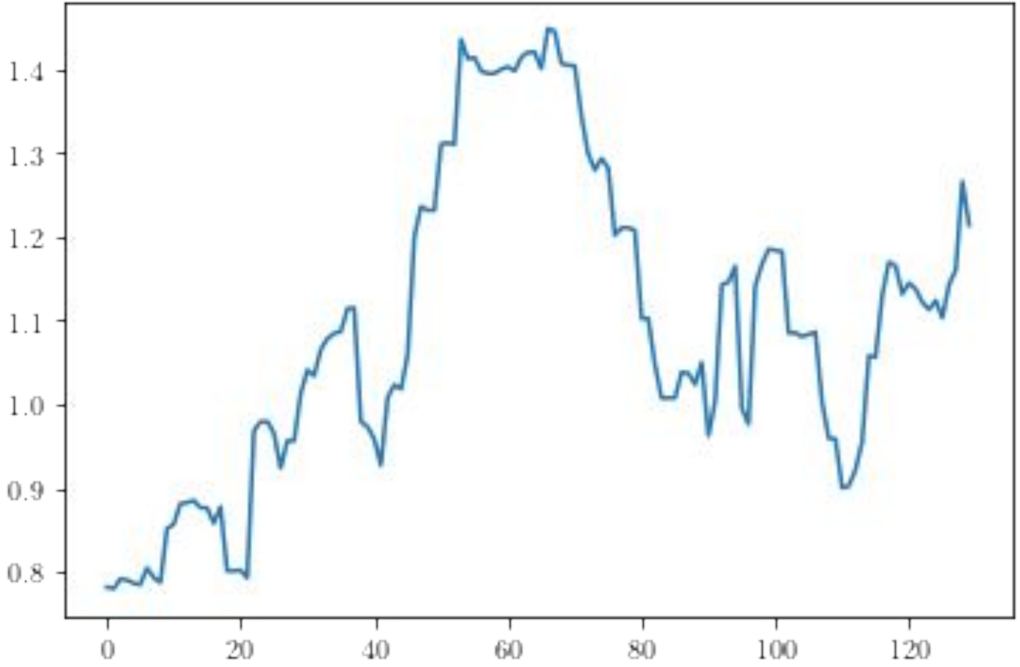




O Captain, My Captain,by Walt Whitman



# O Captain, My Captain!

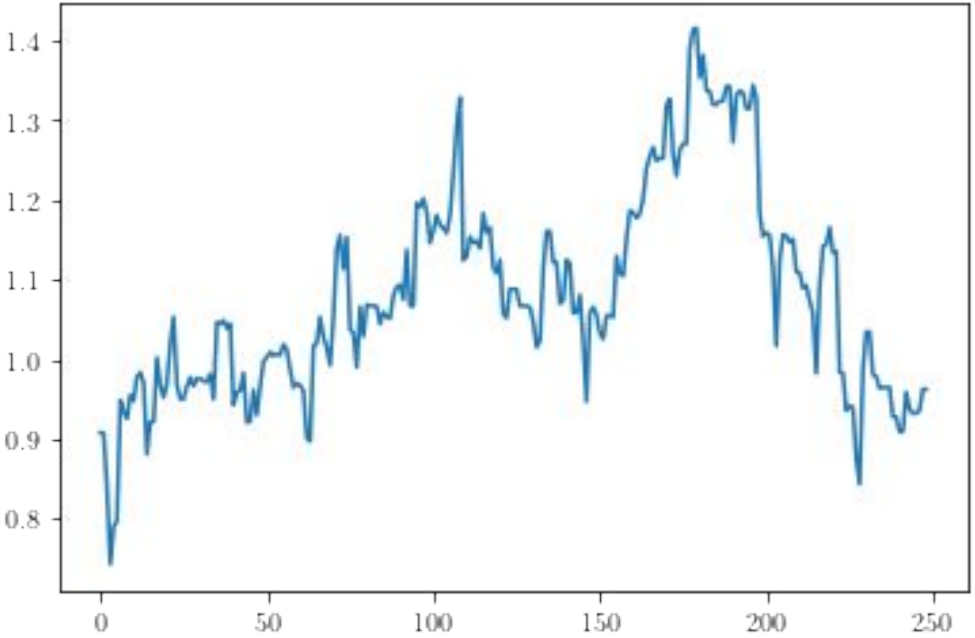


O Captain! my Captain! our fearful trip is done,
The ship has weather'd every rack, the prize we sought is won,
The port is near, the bells I hear, the people all exulting,
While follow eyes the steady keel, the vessel grim and daring;
But O heart! heart! heart!
O the bleeding drops of red,
Where on the deck my Captain lies,
Fallen cold and dead.

O Captain! my Captain! rise up and hear the bells;
Rise up—for you the flag is flung—for you the bugle trills,
For you bouquets and ribbon'd wreaths—for you the shores
a-crowding,
For you they call, the swaying mass, their eager faces turning;
Here Captain! dear father!
This arm beneath your head!
It is some dream that on the deck,
You've fallen cold and dead.

My Captain does not answer, his lips are pale and still,
My father does not feel my arm, he has no pulse nor will,
The ship is anchor'd safe and sound, its voyage closed and done,
From fearful trip the victor ship comes in with object won;
Exult O shores, and ring O bells!
But I with mournful tread,
Walk the deck my Captain lies,
Fallen cold and dead.

# Raven

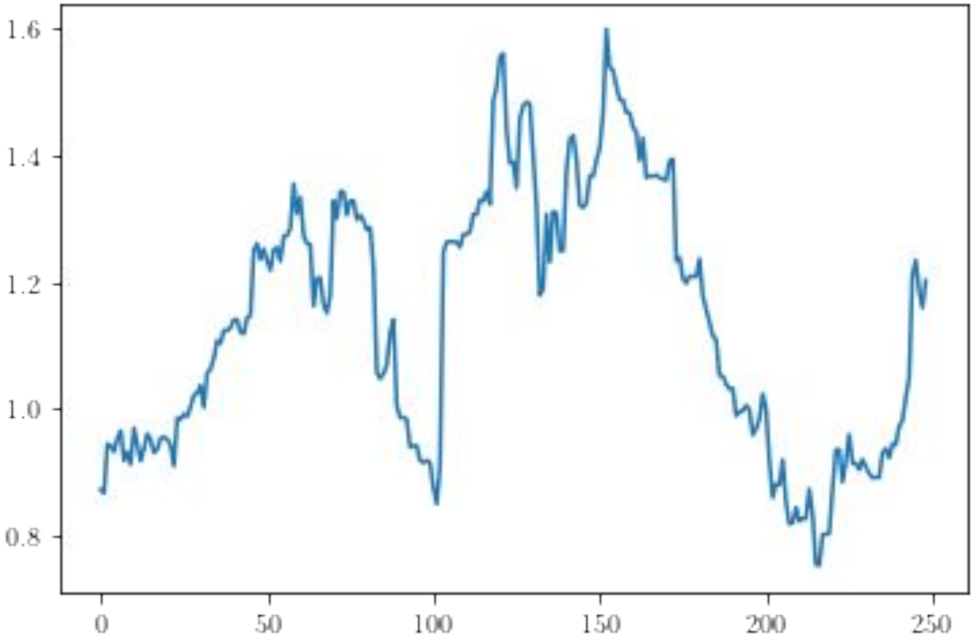


Once upon a midnight dreary, while I pondered, weak and weary,
Over many a quaint and curious volume of forgotten lore—
While I nodded, nearly napping, suddenly there came a tapping,
As of someone gently rapping, rapping at my chamber door.
"'Tis some visitor," I muttered, "tapping at my chamber door—
Only this and nothing more."

Ah, distinctly I remember it was in the bleak December;
And each separate dying ember wrought its ghost upon the floor.
Eagerly I wished the morrow;—vainly I had sought to borrow
From my books surcease of sorrow—sorrow for the lost
Lenore—
For the rare and radiant maiden whom the angels name Lenore—
Nameless here for evermore.

And the silken, sad, uncertain rustling of each purple curtain
Thrilled me—filled me with fantastic terrors never felt before;
So that now, to still the beating of my heart, I stood repeating
"'Tis some visitor entreating entrance at my chamber door—
Some late visitor entreating entrance at my chamber door;—
This it is and nothing more."

# Howl



I saw the best minds of my generation destroyed by madness, starving hysterical naked,
dragging themselves through the negro streets at dawn looking for an angry fix,
angelheaded hipsters burning for the ancient heavenly connection to the starry dynamo in the machinery of night,
who poverty and tatters and hollow-eyed and high sat up smoking in the supernatural darkness of cold-water flats floating across the tops of cities contemplating jazz,
who bared their brains to Heaven under the El and saw Mohammedan angels staggering on tenement roofs illuminated,
who passed through universities with radiant cool eyes hallucinating Arkansas and Blake-light tragedy among the scholars of war,
who were expelled from the academies for crazy & publishing obscene odes on the windows of the skull,
who cowered in unshaven rooms in underwear, burning their money in wastebaskets and listening to the Terror through the wall,
who got busted in their pubic beards returning through Laredo with a belt of marijuana for New York,
who ate fire in paint hotels or drank turpentine in Paradise Alley, death, or purgatoried their torsos night after night
with dreams, with drugs, with waking nightmares, alcohol and cock and endless balls,
incomparable blind streets of shuddering cloud and lightning in the mind leaping toward poles of Canada & Paterson, illuminating all the motionless world of Time between,
Peyote solidities of halls, backyard green tree cemetery dawns, wine drunkenness over the rooftops, storefront boroughs of teahead joyride neon blinking traffic light, sun and moon and tree vibrations in the roaring winter dusks of Brooklyn, ashcan rantings and kind king light of mind,
who chained themselves to subways for the endless ride from Battery to holy Bronx on benzedrine until the noise of wheels and children brought them down shuddering mouth-wracked and battered bleak of brain all drained of brilliance in the drear light of Zoo,
...

# Discussion

There are so many more questions…

- Contextual word embeddings like BERT?
- More “advanced” neural net architecture like LSTM (sequential) or transformer (attention-based)?
- TDA for creativity classification? Plot characterization? Neural network interpretability?

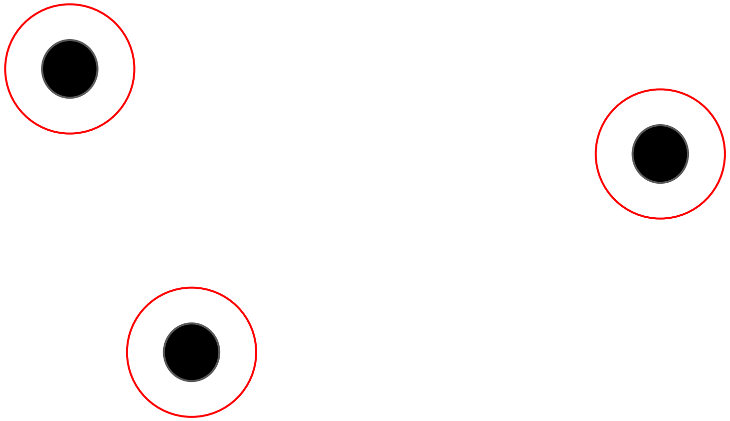
Thank you! :)

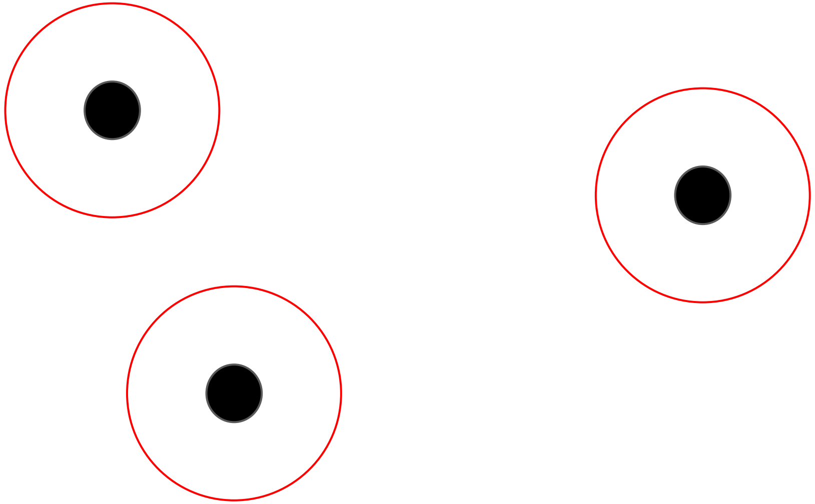
# Conclusion

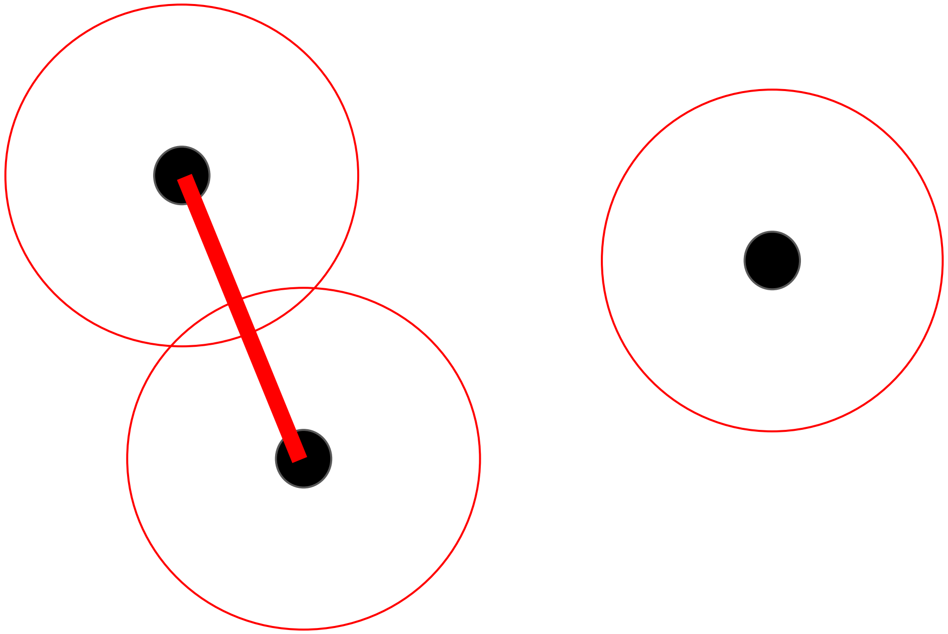
Topological perspective on literature opens us a world of opportunity

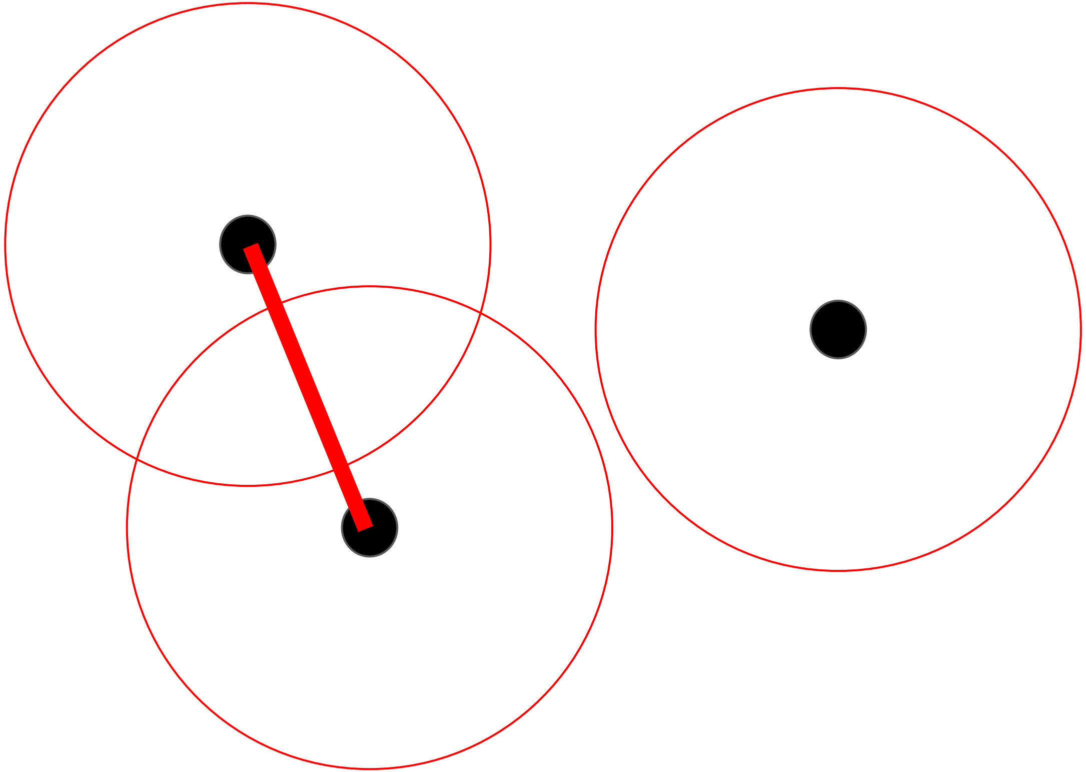
- Text classification (authorial signature)
- Plot analysis (change-point detection!)
- Text summarization

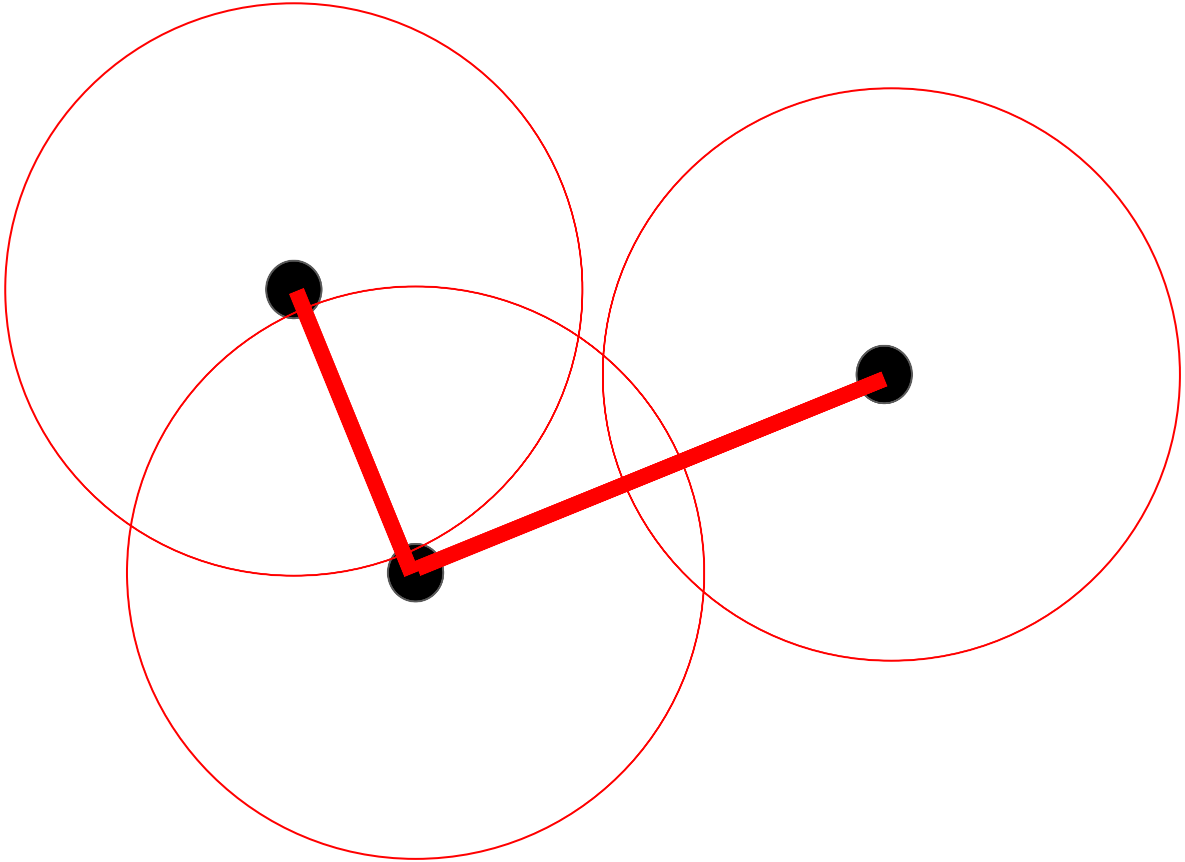
Preserves understanding of the sequential nature of words (unlike many NLP problems which treat text as a “bag of words”)

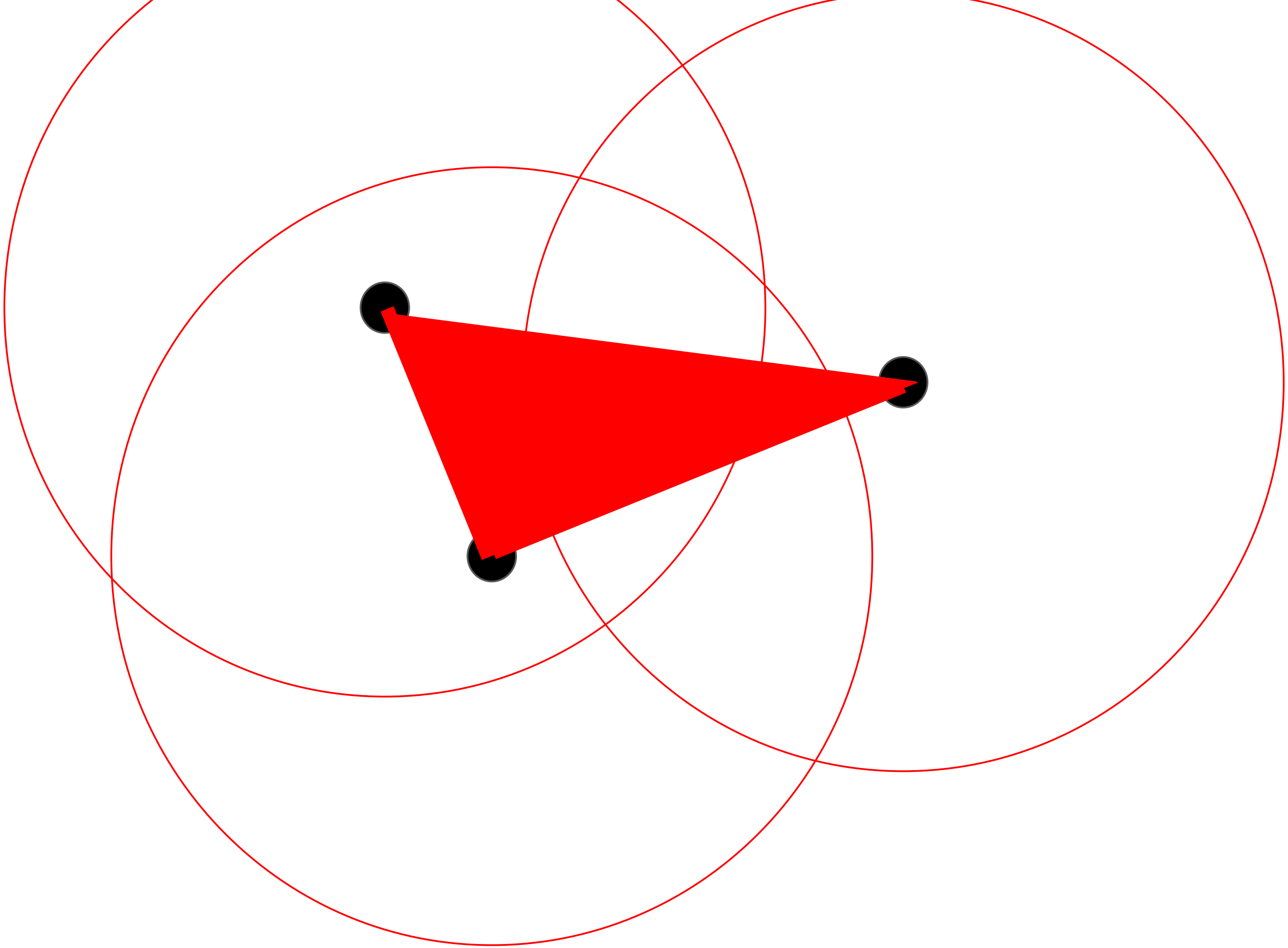


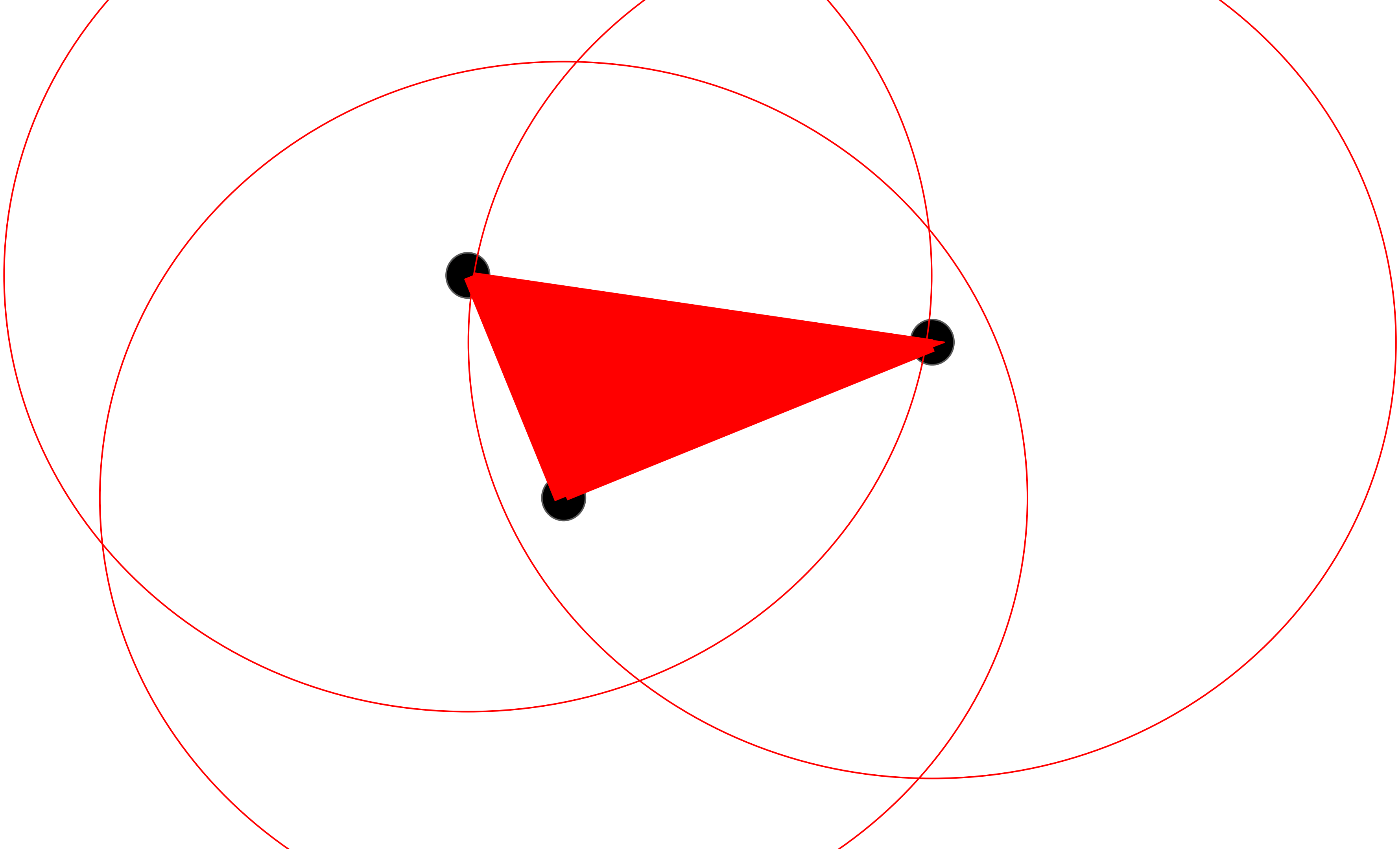






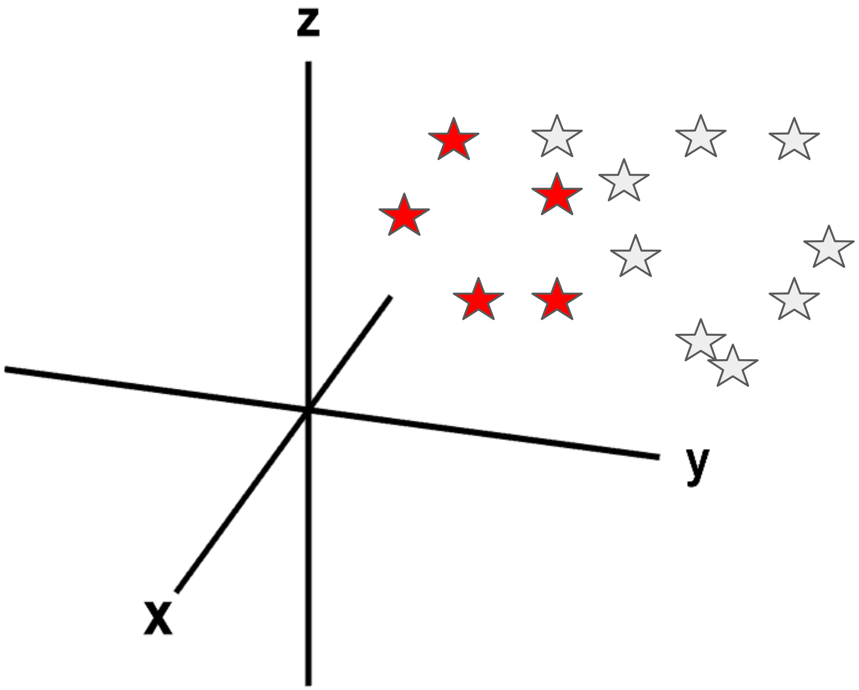






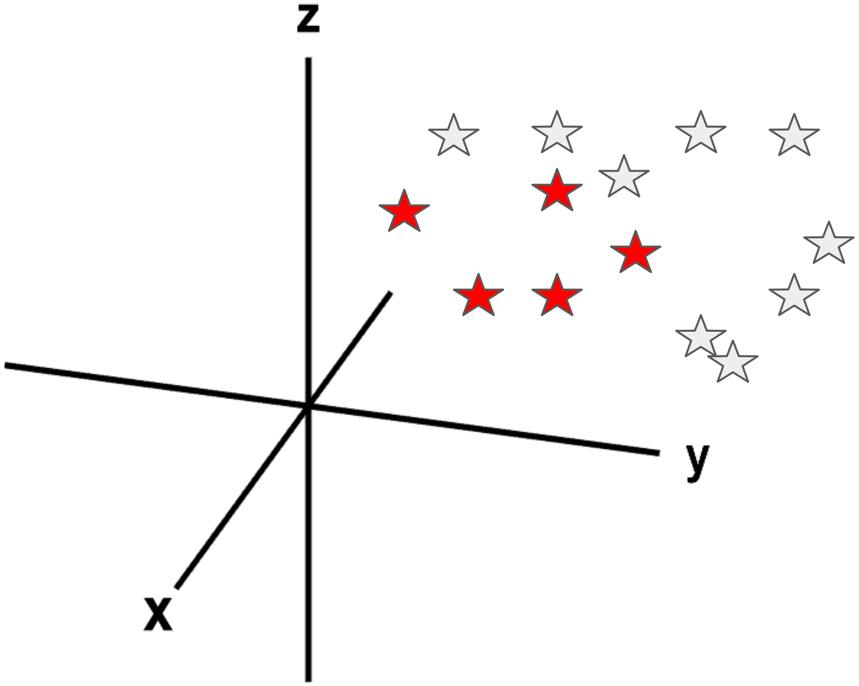
POINTS → A, B, C, D, E, F, G, H, I, J, …

Run TDA on this window (persistence norm!)



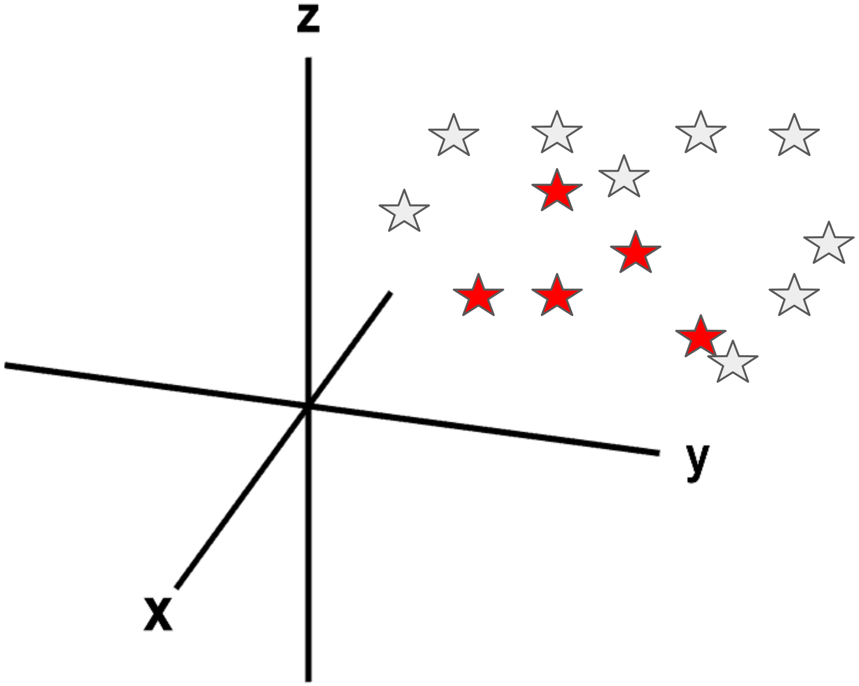
POINTS → A, B, C, D, E, F, G, H, I, J, …

Run TDA on this window (persistence norm!)



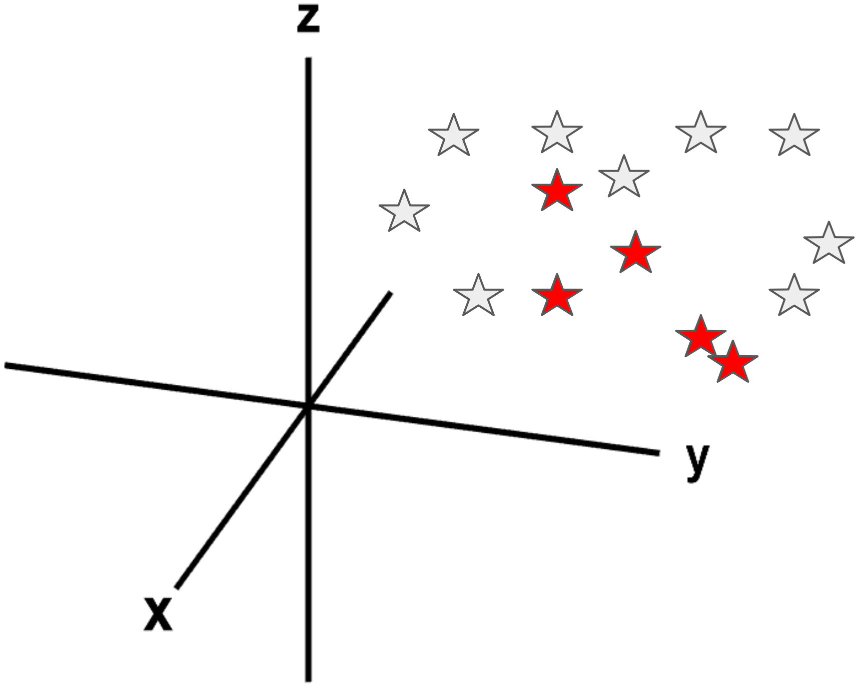
POINTS → A, B, C, D, E, F, G, H, I, J, …

Run TDA on this window (persistence norm!)



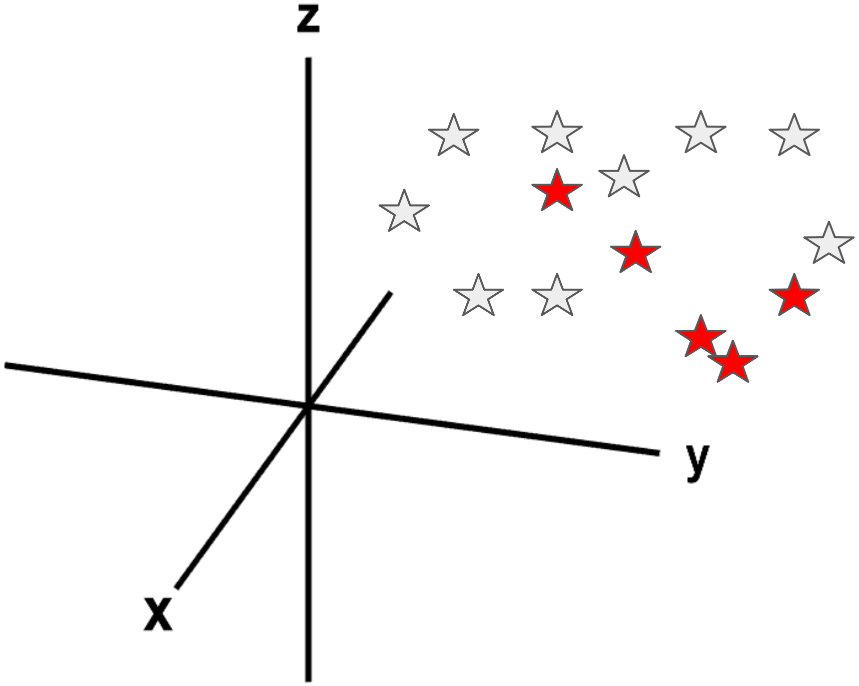
POINTS → A, B, C, D, E, F, G, H, I, J, …

Run TDA on this window (persistence norm!)



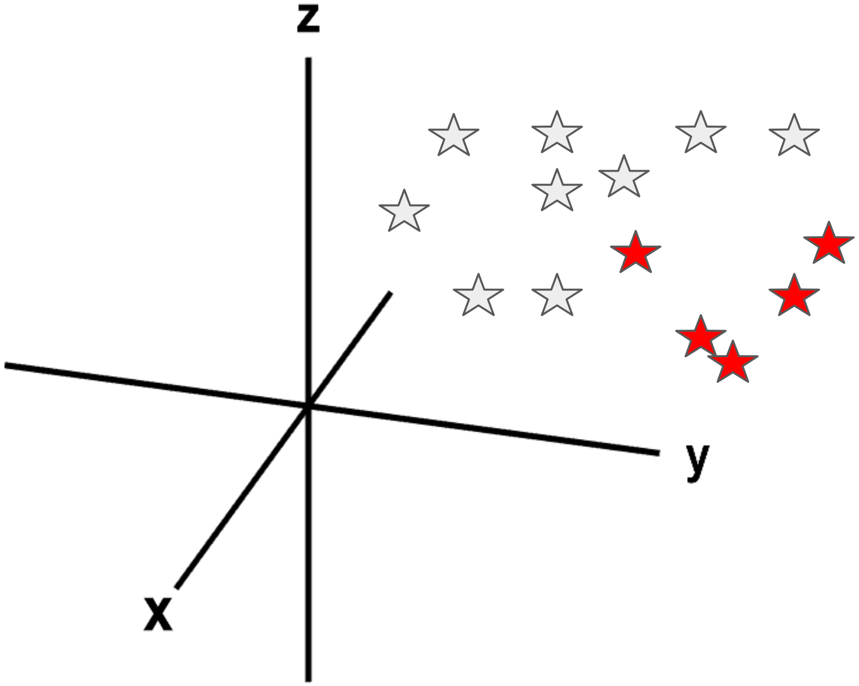
POINTS → A, B, C, D, E, F, G, H, I, J, …

Run TDA on this window (persistence norm!)



POINTS → A, B, C, D, E, F, G, H, I, J, …

Run TDA on this window (persistence norm!)



# Abstract

Topological data analysis (TDA), in contrast with more traditional statistical methods, allows us to quantify the shape and effectively reduce the dimensionality of high-dimensional, nonlinear data. In this project, we use TDA to enhance a simple linear regression model aimed at automated author classification. Motivated by previous applications of TDA to time series data, we collect topological metrics of the sequence of word embeddings and use these as features. We find that these TDA-driven metrics outdo more direct statistical analyses of the word embeddings such as max or sum pooling. This project sheds light into benefits of incorporating TDA into the natural language processing and general machine learning pipeline.

# Previous Work

Zhu (2013), which is considered the first application of persistent homology to NLP, developed the the Similarity Filtration with Time Skeleton (SIFTS) algorithm, which provided a new document structure representation

Gholizadeh (2018) analyzed authorial signatu [redacted] main characters in the novel.

Elyasi (2019) used word embedding on a ver renowned Iranian poets.

# Columbia University Undergraduate Mathematics Research Symposium

April 20, 2022

COLUMBIA UNIVERSITY

IN LVMINE TVO VIDEBIMVS LVMEN