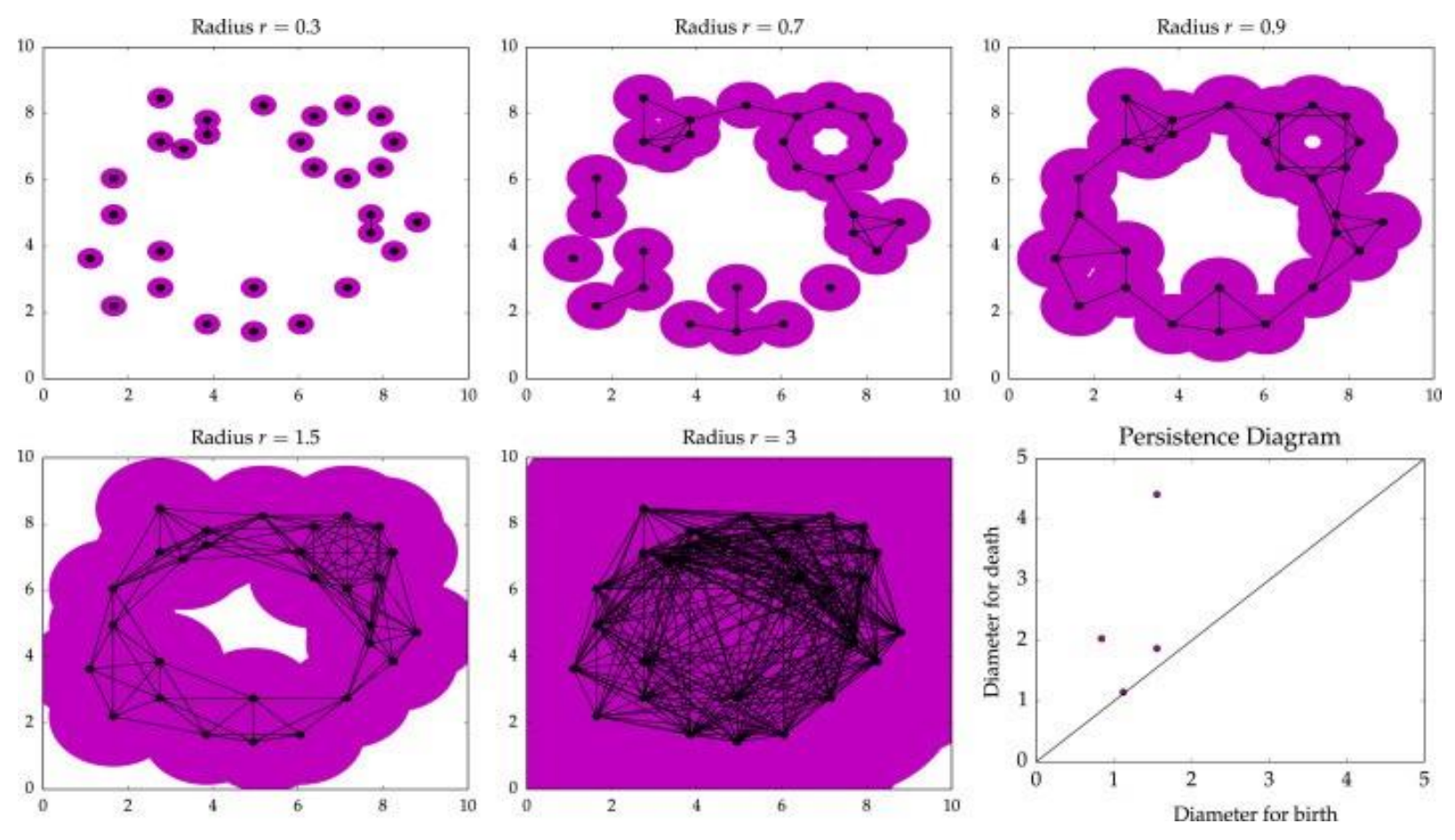
TOPOLOGICAL DATA ANALYSIS FOR WORD SENSE DISAMBIGUATION

Michael Rawson, Mithun Bharadwaj ,Samuel Dooley, Rishabh Choudhary

What is TDA and Persistent Homology

- The purpose of topological data analysis is to apply the tools of **topology**, a field of mathematics dealing with qualitative geometric features such as smoothness and connectedness to analyze datasets.



- **Persistent homology** is a method for computing topological features of a space at different spatial resolutions.
- For example, we decide on a radius r = 0.5. And for each point in our dataset, we connect that point with lines to all other points which are within this radius from it.
- We gradually and arbitrarily increase this radius. So that more and more points get connected to each point. To the extreme situation when the radius is so big that all points get connected to all points and our entire space of data points are covered and connected to each other.

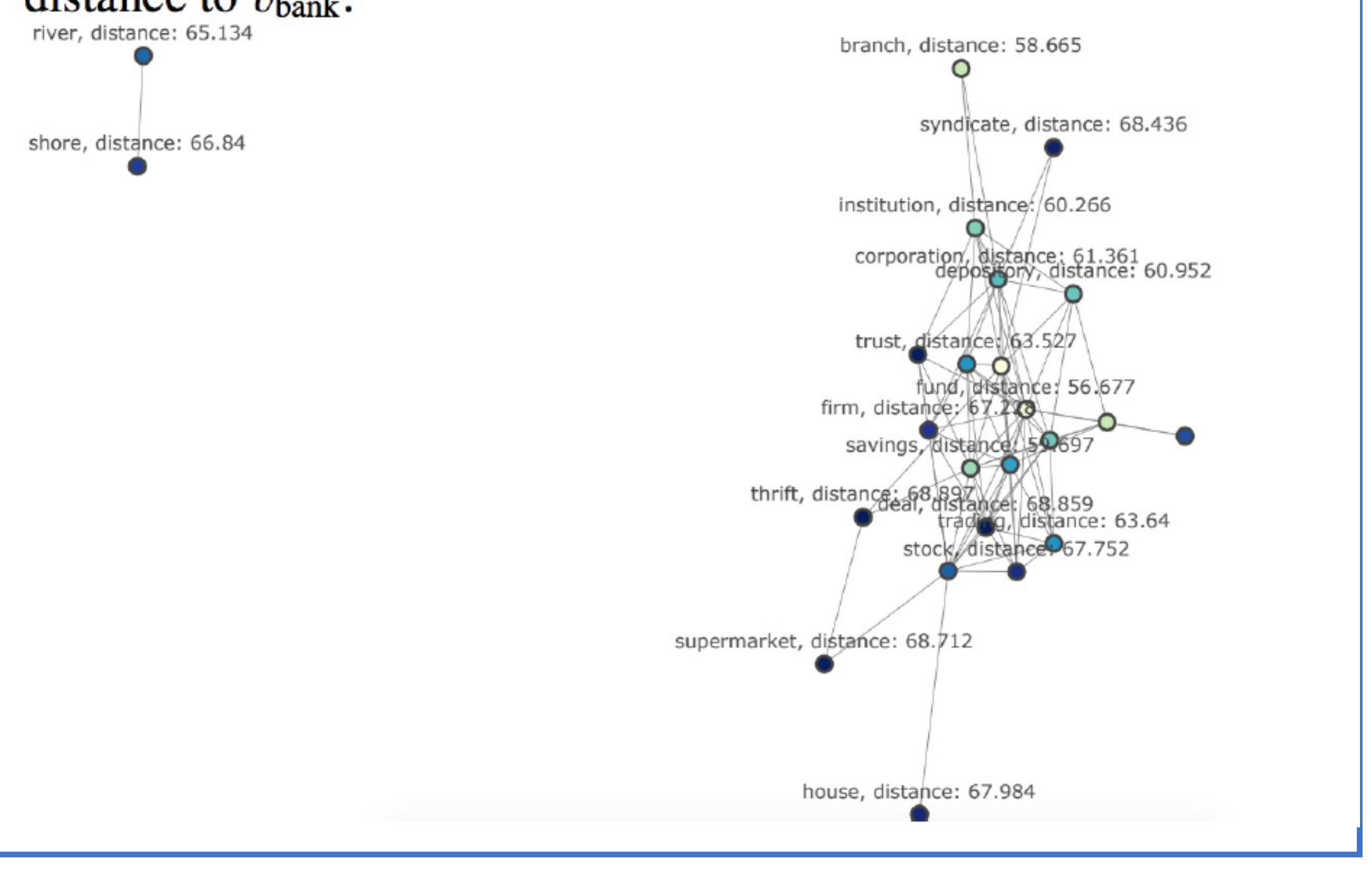
Research Questions

- Can topological data analysis of word embeddings sufficiently recover word senses?
- How can persistent homology be extended to NLP to disambiguate the different senses of a word?
- Is the topology of word embeddings for WSD tasks or its usage in NLP justified since it's a relatively new application in this realm?
- How do unsupervised clustering algorithms to determine word senses generally fare?
- How does the result of the clustering algorithm vary with the number of closest words considered to predict the number of senses?

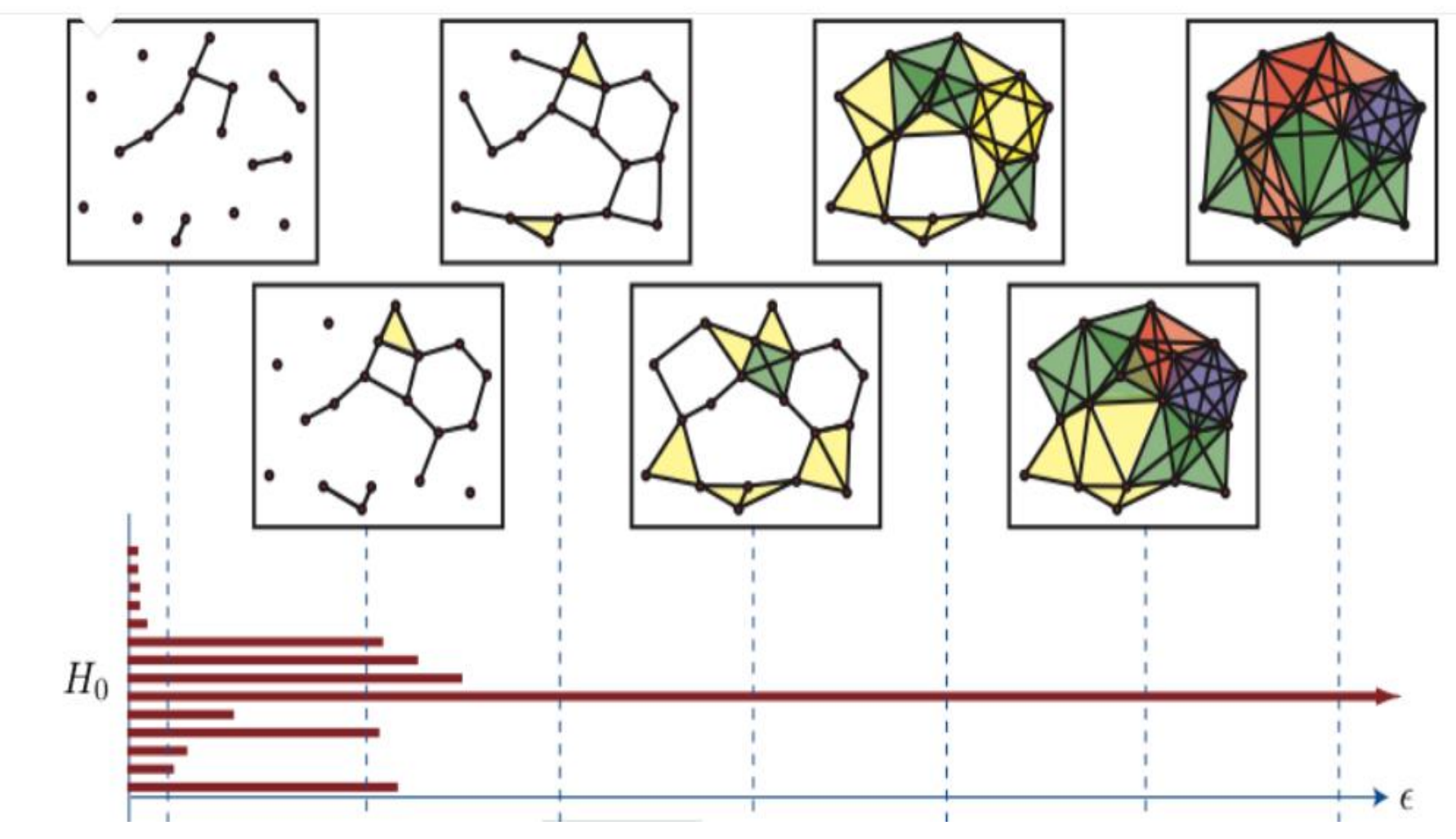
Problem and motivation: Why TDA?

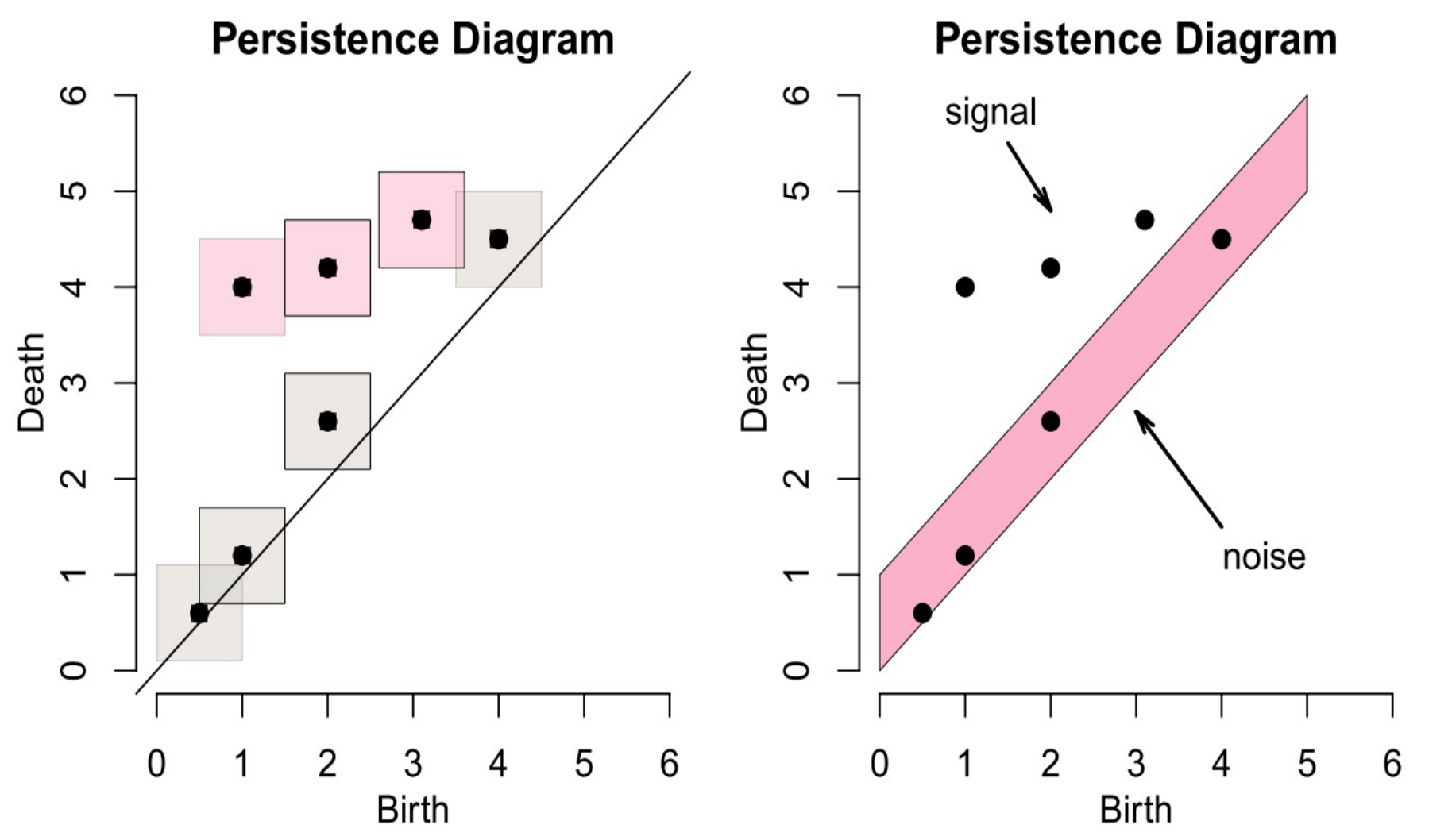
- Often, we do not know what the shape of the data is, but believe the data has been sampled from some underlying surface or manifold lesser in dimension than the ambient space.
- Extraction of information from datasets that are high-dimensional, incomplete and noisy is generally challenging. TDA provides a general framework to analyze such data in a manner that is insensitive to the particular metric chosen and provides dimensionality reduction and robustness to noise.
- Stratification is a decomposition of a topological space into manifold like pieces. When thinking about stratification learning and word embeddings, it seems intuitive that vectors of words corresponding to the same broad topic would constitute a structure, which we might hope to be a manifold.
- Hence, for example, by looking at the intersections between those manifolds one might hope to algorithmically find vectors of homonyms like 'bank' (which can mean either a river bank, or a financial institution) or vectors of words with very different meanings like 'cancer' (which can refer to the Cancer constellation or to the illness).

Figure 11. The link of v_{bank} at $\epsilon = 69°$; distance is the geodesic distance to v_{bank}.



Technical Approach: Barcode Algorithm





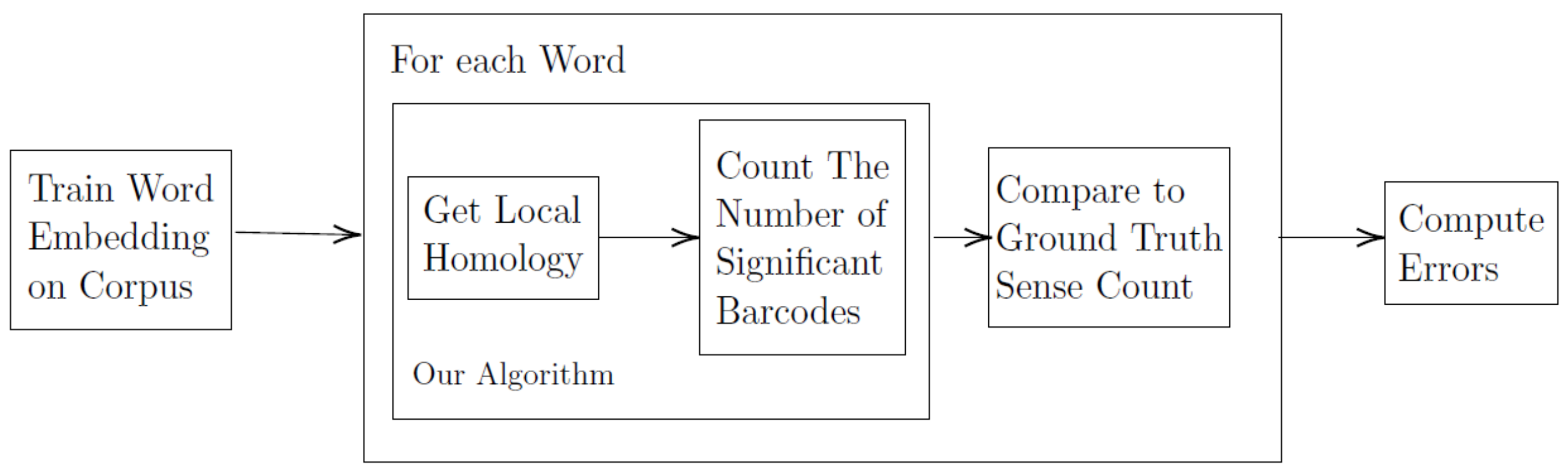
- We explain the barcode algorithm for the 0th persistent homology. We start with a list of N data points, X, each in d dimensional Euclidean space.
- If we increase the radius and plot the radius along the horizontal x-axis. And for each value of the radius, you connect the points in your dataset to create a simplicial complex, you will see various features (e.g. holes) appearing and disappearing in your simplicial complex.
- We plot these features as bars (barcodes) along the horizontal axis. The short bars correspond to noise and the long bars correspond to features.
- After removing the noise, the number of connected components will correspond to the different senses of the word.

Evaluation and Results

Data:

- Our baseline results have been computed on the Semcor dataset annotated using WordNet with the word2vec embedding training routine.

Evaluation pipeline:



Train Word Embedding on Corpus
For each Word
Get Local Homology
Count The Number of Significant Barcodes
Our Algorithm
Compare to Ground Truth Sense Count
Compute Errors

- We define (**number of senses = number of death dates that are > (mean + 2*std))**.
- We compute absolute and relative errors when compared to the ground truth number of senses.

Results:

2 ≤ sense_count_dict[word] ≤ 9

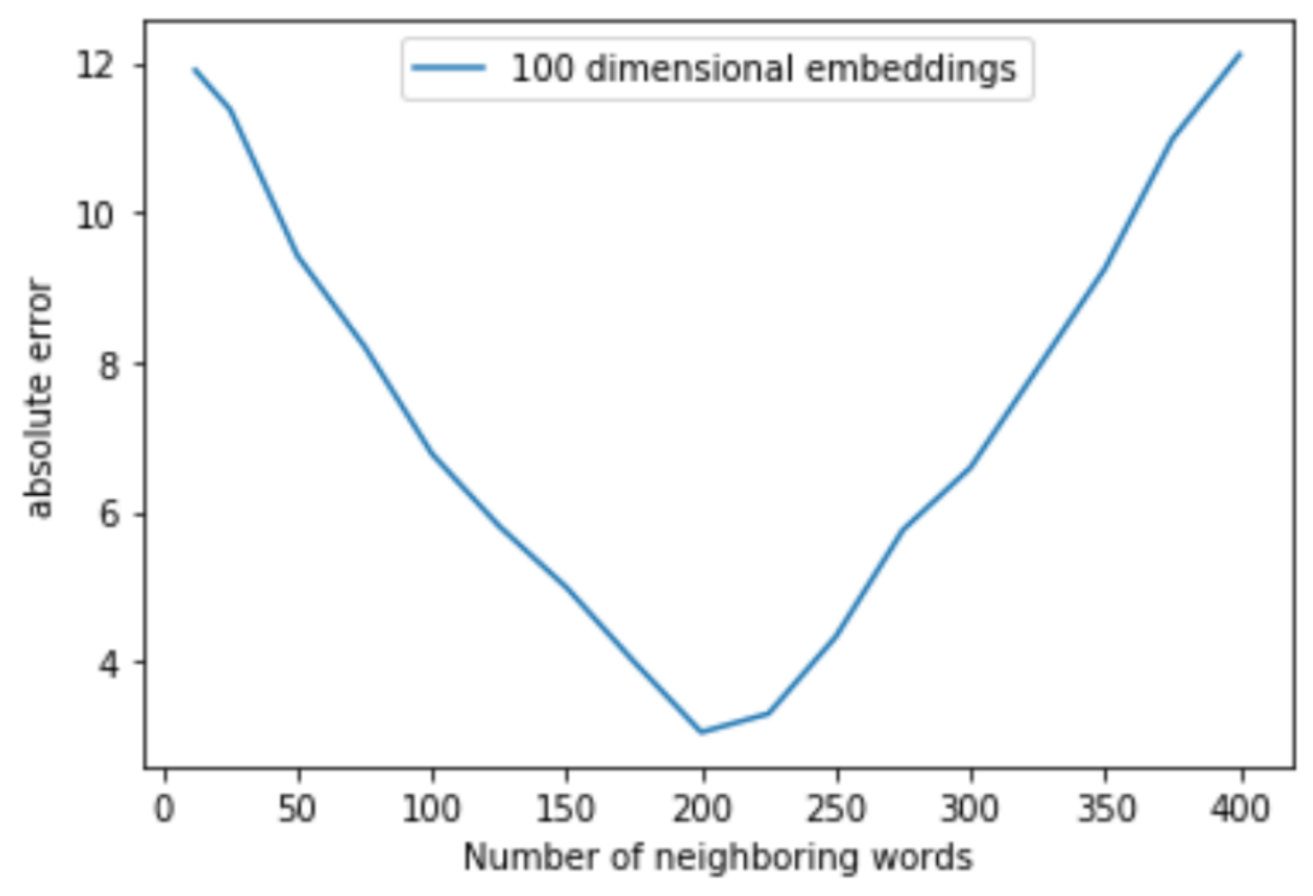
Embedding Dimension	Num Neighbors Considered	Relative Error	Absolute Error
500	100	0.9023	2.723
500	50	0.4178	1.660
500	25	0.4114	2.202
100	100	0.8030	2.447
100	50	0.4331	1.745
100	25	0.4348	2.287

10 ≤ sense_count_dict[word] ≤ 19

Embedding Dimension	Num Neighbors Considered	Relative Error	Absolute Error
1000	200	0.2129	2.907
1000	100	0.4609	6.511
500	200	0.1897	2.6511
500	100	0.4712	6.6279
100	200	0.2068	3.0465
100	100	0.4855	6.5116

Analysis: Graph of number of closest words considered vs absolute error

- If we consider very few neighboring words, we have less information than what is required and for a large number words we have too much irrelevant information. In both cases error is high as expected.
- The sweet spot is in the middle, hence our results are coherent with the intuitive expectation



Future Work / Additional Ideas

- How will the clustering algorithm be affected by different word embeddings, i.e if GloVe or FastText is used instead of Word2Vec?
- How does the algorithm perform with pseudowords? More specifically for a word that has multiple meanings or senses, we can break that word up into its component senses, create new dummy words, and replace the original word with the respective dummy word. For example, if the word "foo" has two meanings, we can create words "foo$1" and "foo$2" and replace "foo" with the respective pseduoword we just created. Then evaluate our algorithm on the created data. This can be a second way to evaluate our algorithm.
- How does the current algorithm compare to other supervised and unsupervised WSD algorithms.

Key References:

- Tadas Temcinas. Local homology of word embeddings. (2018)
- Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. - Confidence sets for persistence diagrams. (2014)
- Niyogi, P., Smale, S., and Weinberger, S. - Finding the homology of submanifolds with high confidence from random samples. (2008)