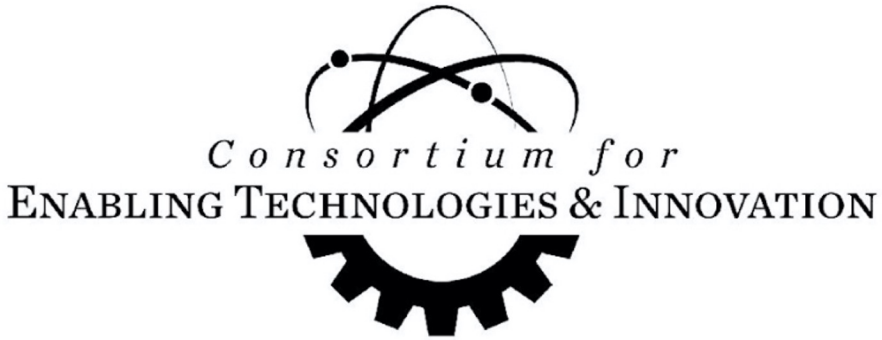
A Geometry-Driven Approach to Longitudinal Topic Modeling of Nuclear-Scientific Literature

Conrad D. Hougen, Karl T. Pazdernik, Alfred O. Hero

Date 03/30/2022



Project Objective

- Primary Objective: Identify potential repurposing of nuclear-scientific research
  - ETI Thrust 1: Computer & Engineering Sciences for Nonproliferation

- Hypothesis: Nuclear science literature contains latent information which can allow for early threat detection
  - Emergence, convergence, and divergence of topics
  - Author collaborations and interests
  - Anomaly detection problem

Open Questions

- How do research topics evolve in the nuclear science domain?
- How does information diffusion occur in the nuclear science community?
- Can we characterize anomalous author behavior?

Proposed Approach

- Initial Approach: Dynamic Topic Modeling
  - Model temporal dynamics within nuclear science research
  - Visualize topic trajectories based on manifold learning
  - Extend static LDA to model temporal information diffusion

- Future Work: Predictive Collaboration Network
  - Predict author collaborations based on author interests
  - Author interests can be modeled as a latent generative process
  - Compare predicted collaborations with public co-authorship data

Motivation: Dynamic Topic Model

- Static topic models ignore temporal information
  - Examples: LDA, NMF
  - Document corpora evolve over time
  - Emergence and disappearance of topics is common

- Dynamic topic modeling provides richer interpretation
  - Allows for visualization of topic evolution
  - Temporal topic manifold can be extended for predictive modeling
  - Models real-world dynamics of author collaborations

Prior Works

- Dynamic LDA
  - Blei and Lafferty [2] proposed a dynamic topic modeling approach
  - Topical alignment captured by a Kalman filter procedure

- Geometry-Driven Longitudinal Topic Model
  - Wang and Hougen et. al. [1] proposed a slice-by-slice approach based on PHATE [3] geometric embedding
  - PHATE: Potential of Heat-diffusion for Affinity-based Trajectory Embedding
  - PHATE preserves geometry of high-dimensional time-varying data
  - Any static topic model can be applied to each time slice

[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. (2021). A Geometry-Driven Longitudinal Topic Model. *Harvard Data Science Review,* 3(2). https://doi.org/10.1162/99608f92.b447c07e

[2] Blei, D. M., & Lafferty, J.D. (2006). Dynamic topic models. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113-120). Omni Press.

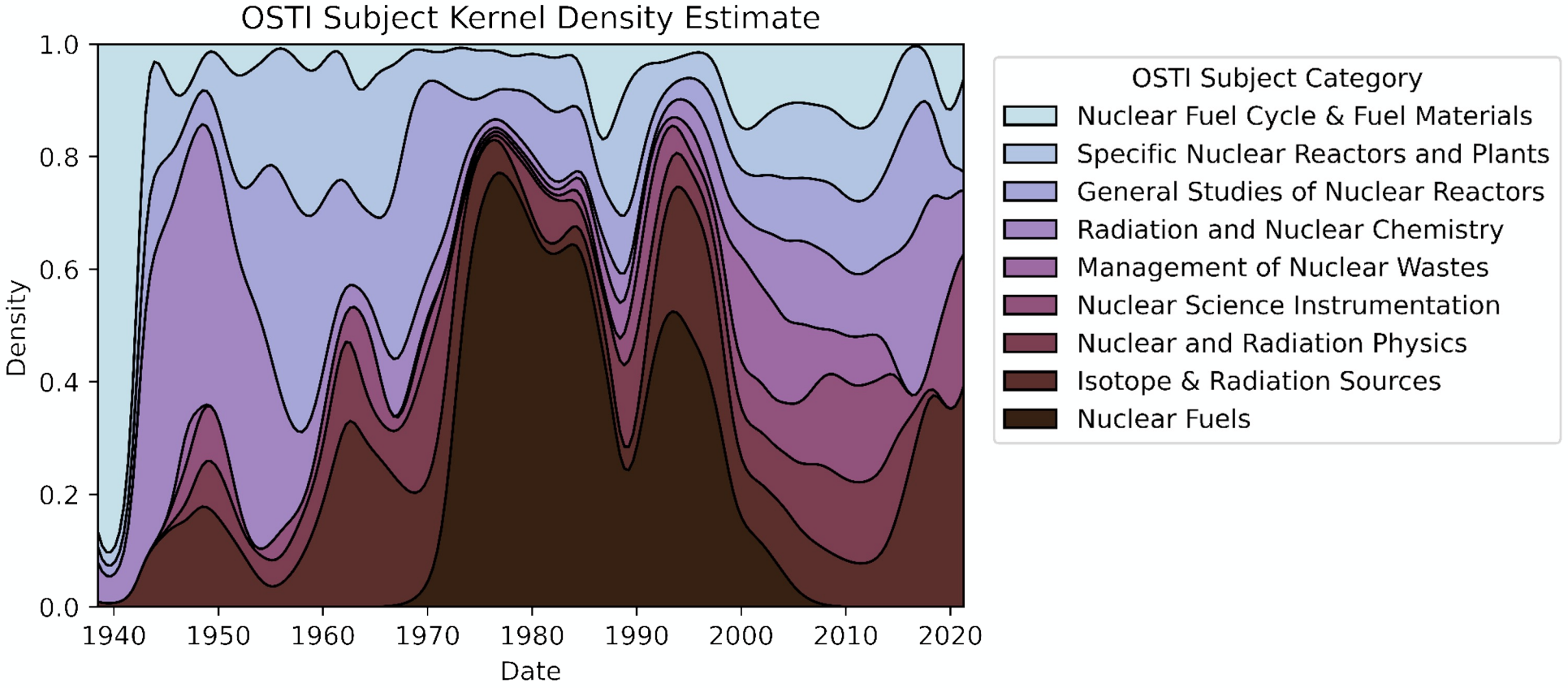
[3] Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, *37* (12), 1482–1492.

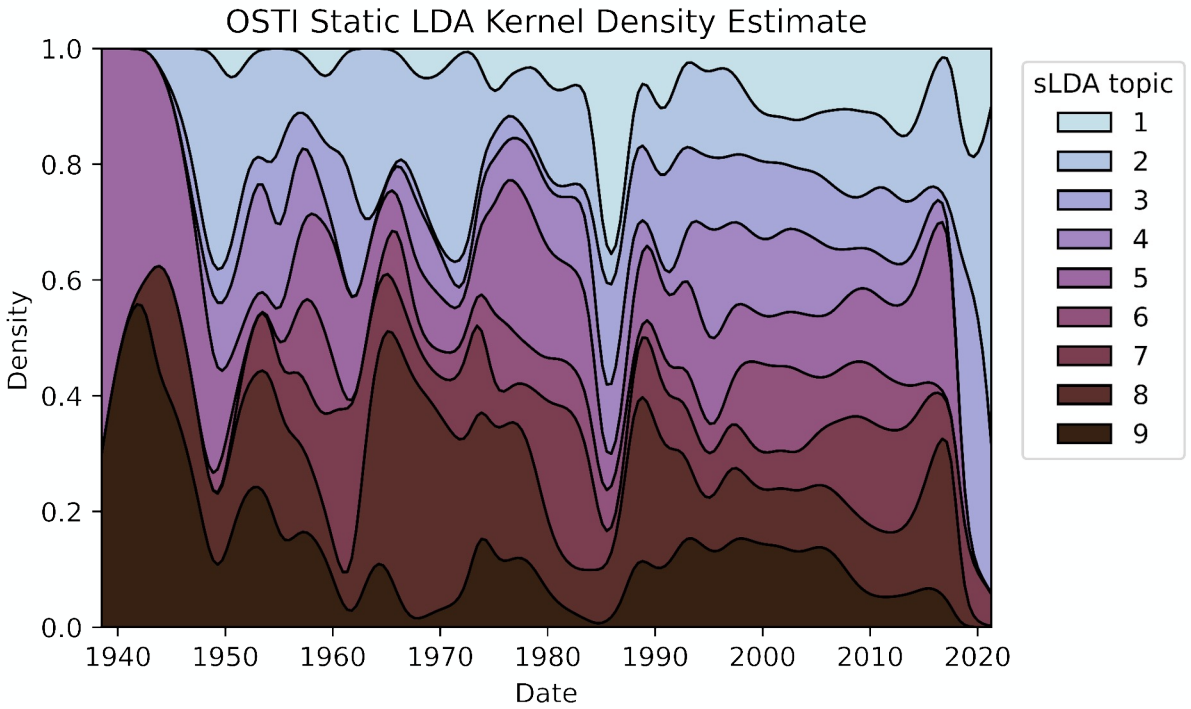
Dataset Exploration

- OSTI Dataset
  - Approx. 200,000 abstracts after cleaning
  - Nuclear-science and unrelated articles
  - OSTI articles are self-labeled by authors
  - 9 OSTI topic labels are associated with the nuclear fuel cycle (NFC)

- Test Dataset
  - Approx. 23,000 NFC-related abstracts
  - All 9 OSTI topic labels represented
  - Data from 1941 - 2018

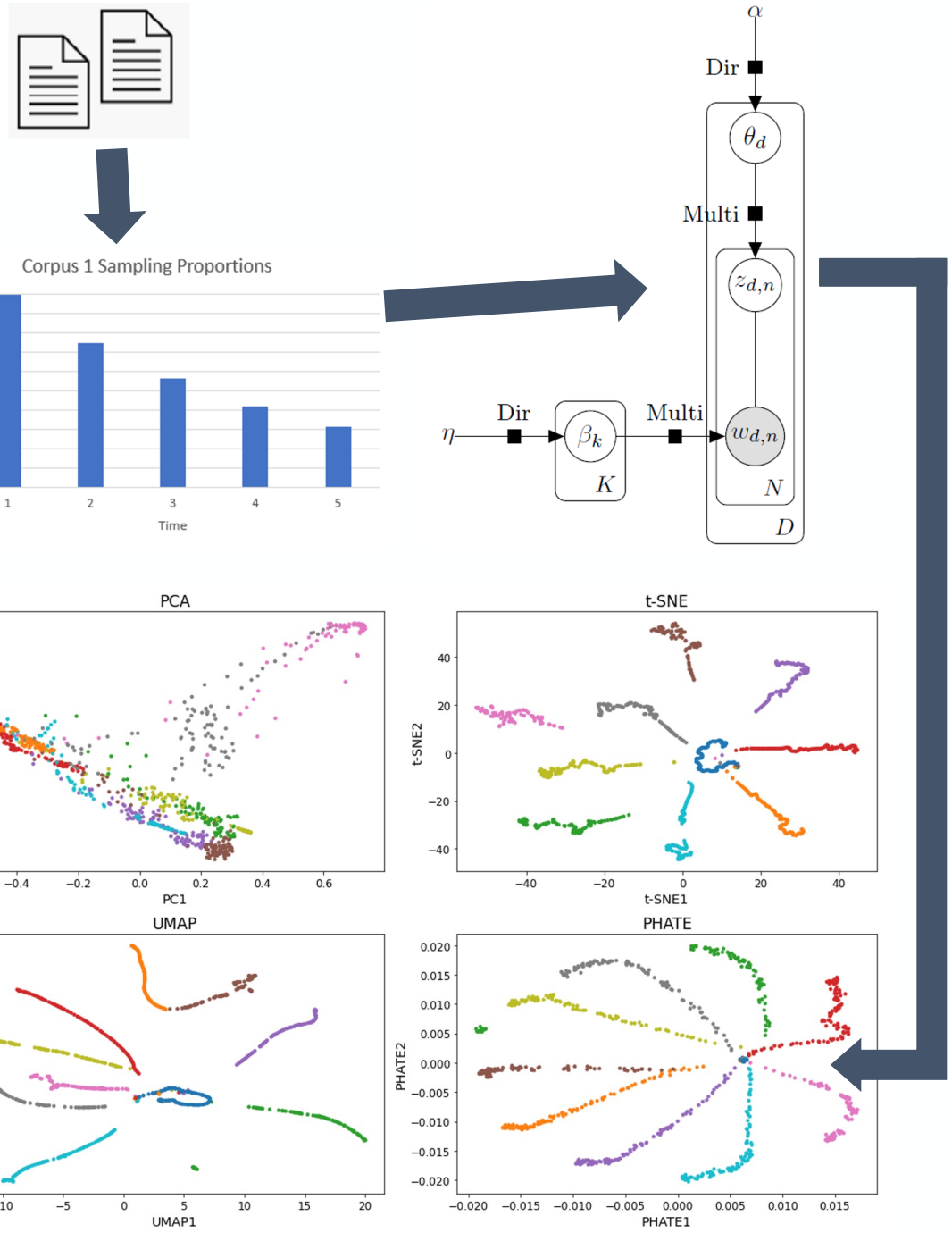
OSTI Dataset: Self-Labeled Topics





Static Topic Model

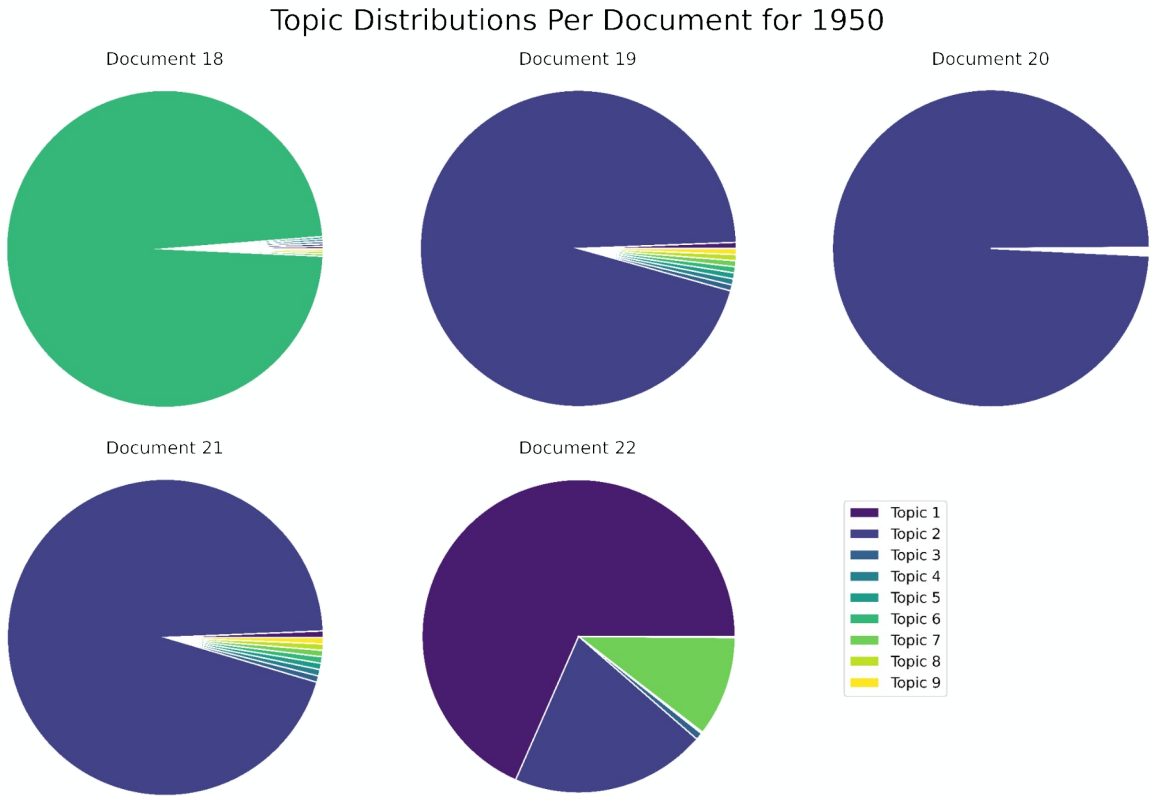
- Latent Dirichlet Allocation (LDA)
  - Apply LDA to entire corpus
  - Assume we want to discover 9 topics
  - LDA assumes documents are generated by latent topics
- Interpretation
  - Can visualize topic word clouds
  - Generate word-frequency and document-topic distributions



Dynamic Topic Model

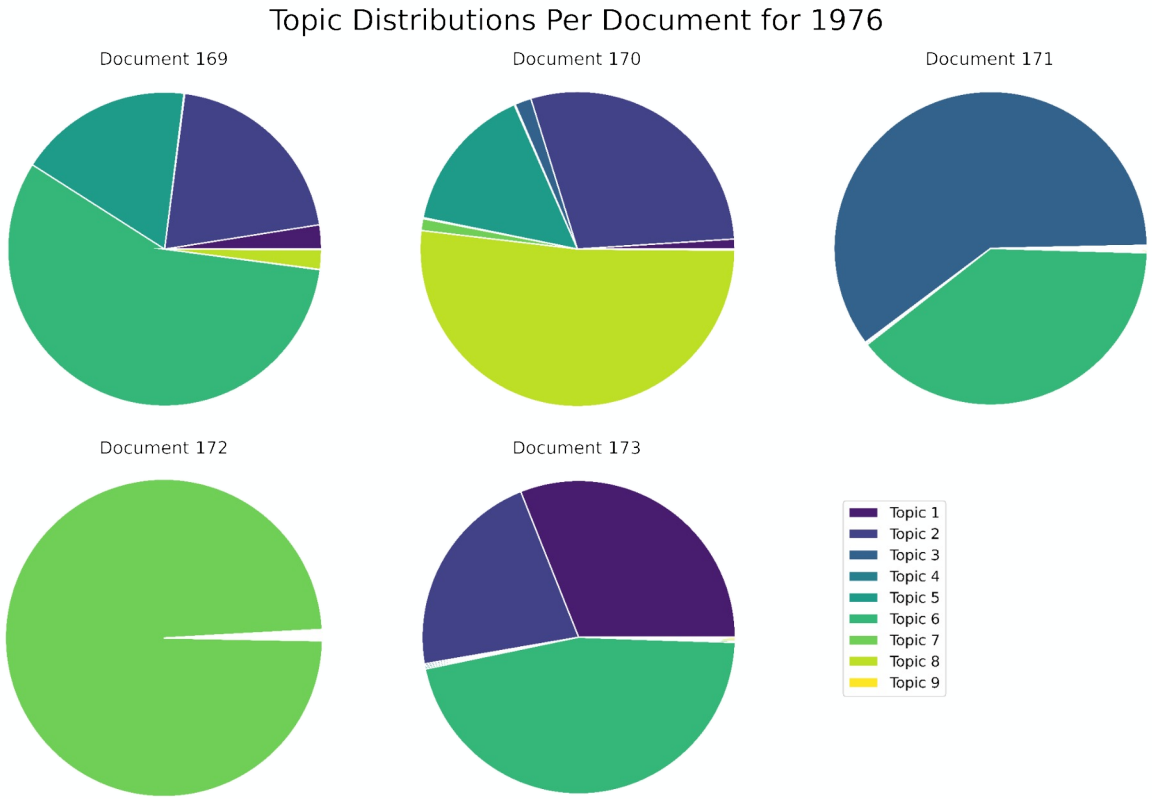
1) Bucket articles by publication date
2) Generate temporally-smoothed corpora
3) Apply LDA "slice-by-slice" i.e. to each sub-corpus
4) Compute distance metric between subtopics (Hellinger distance)
5) Perform hierarchical clustering
6) Apply PHATE low-dimensional embedding

ETI



Documents and Topics

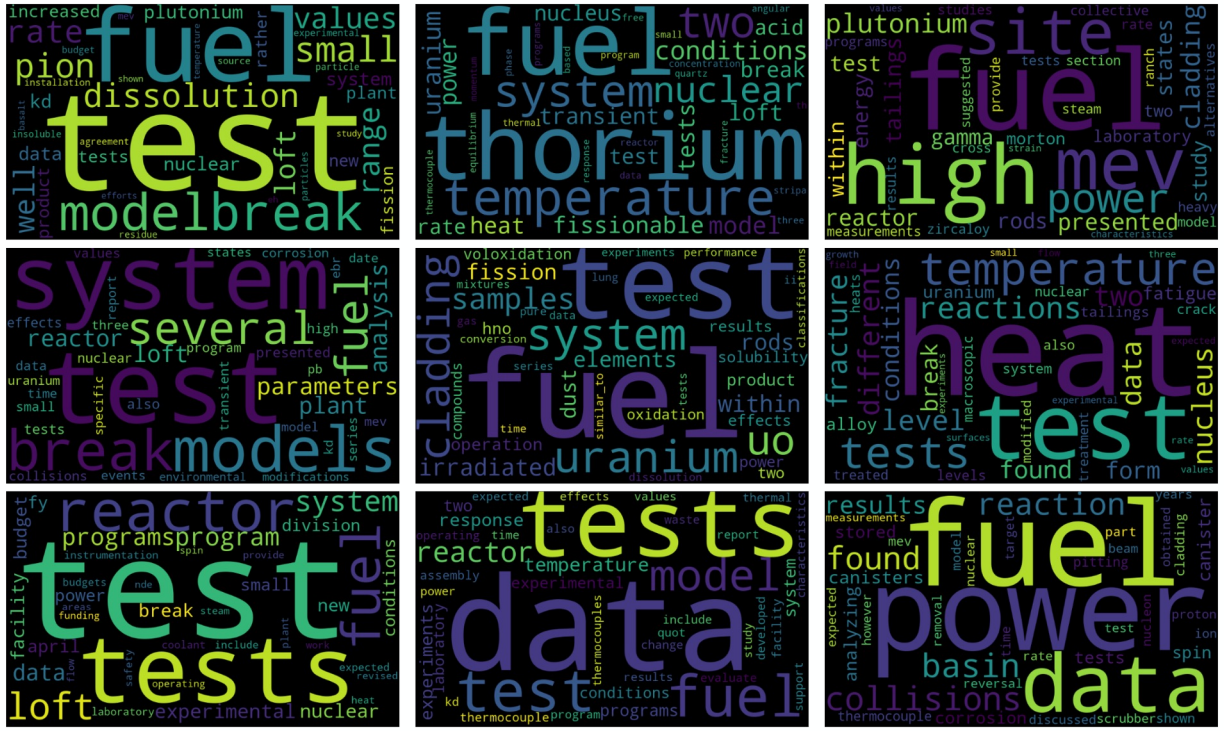
- Documents are formed from several latent topics
- Some documents are dominated by only a few topics
- Latent topics uncovered by LDA may be more representative than OSTI labels



Documents and Topics

- OSTI documents are self-labeled into single categories
- Static LDA on time slice shows split of some document over multiple topics
- What does this mean?
- Can we learn the low-dimensional manifolds on which topics and documents exist?

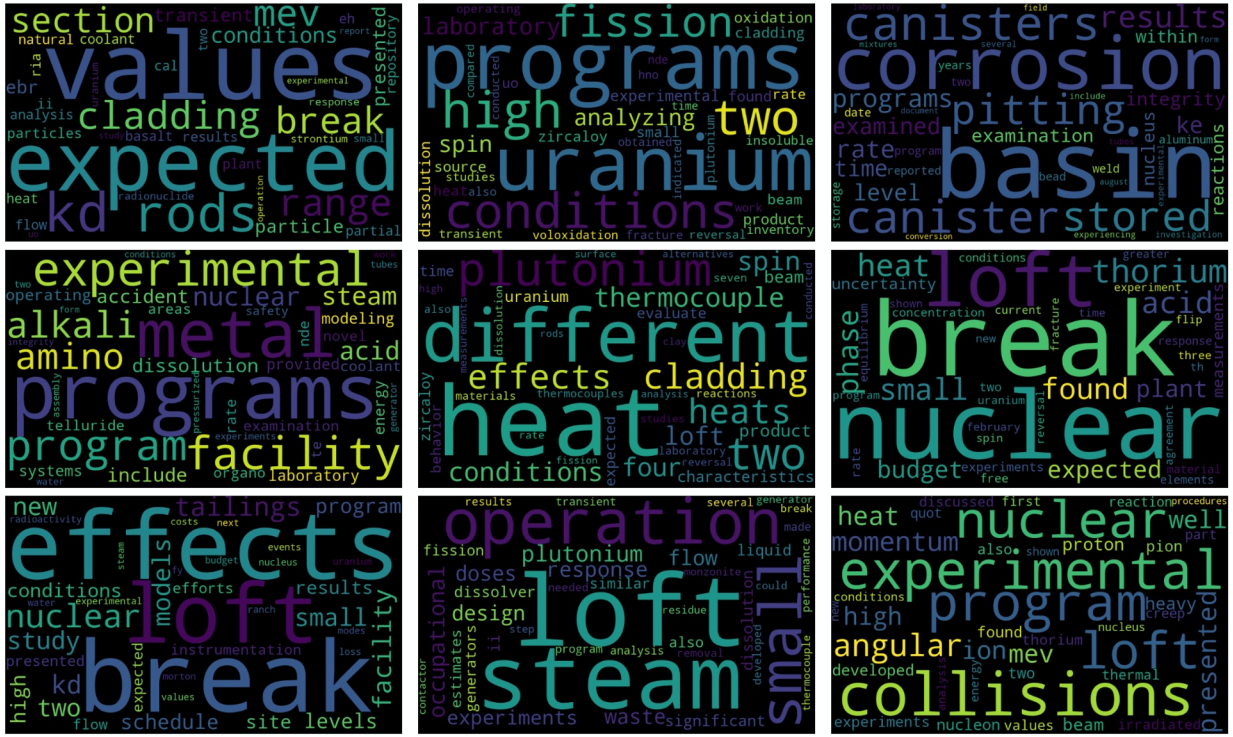
Topic Wordclouds for 1980



Topic Visualization

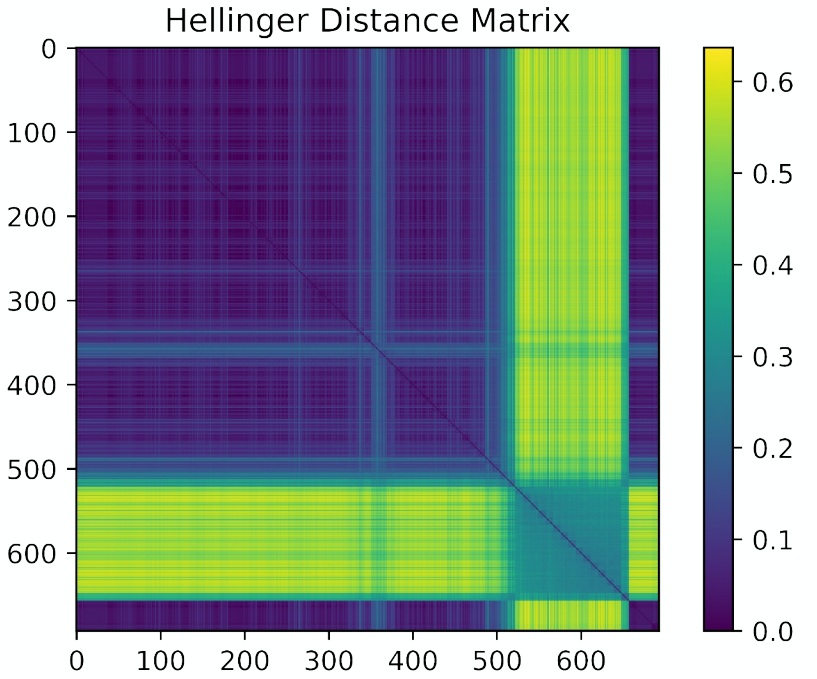
- Word clouds are a common method for understanding documents and topics
- Left: snapshot of LDA applied to 1980 time slice
- Some words are extremely prevalent in literature but not revealing

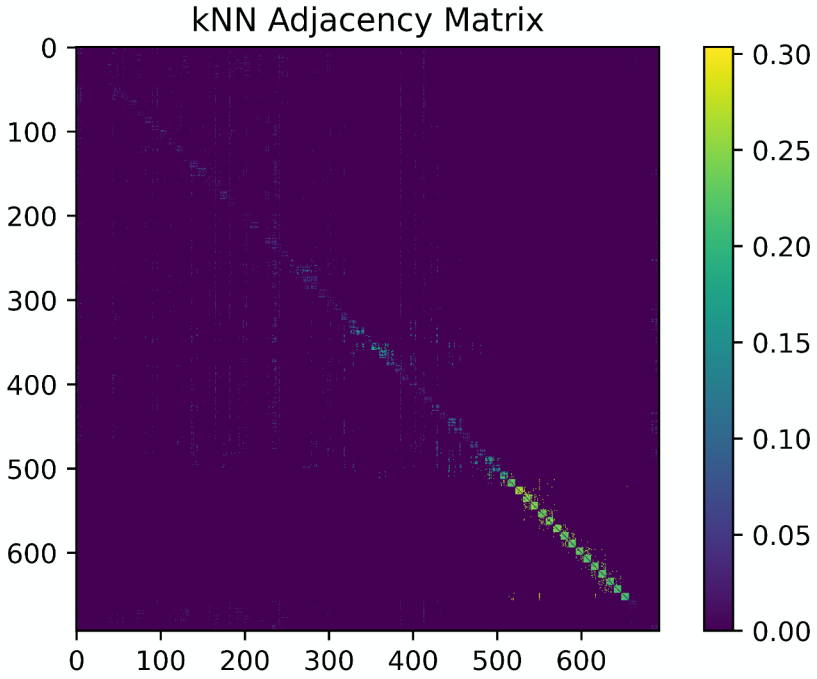
Conditional LDA Topic Wordclouds for 1980



Topic Visualization

- Conditional LDA model can be created based on initial model
- Remove words which are high-frequency in specific domain
- Better differentiation of topics
- Still want to understand temporal evolution

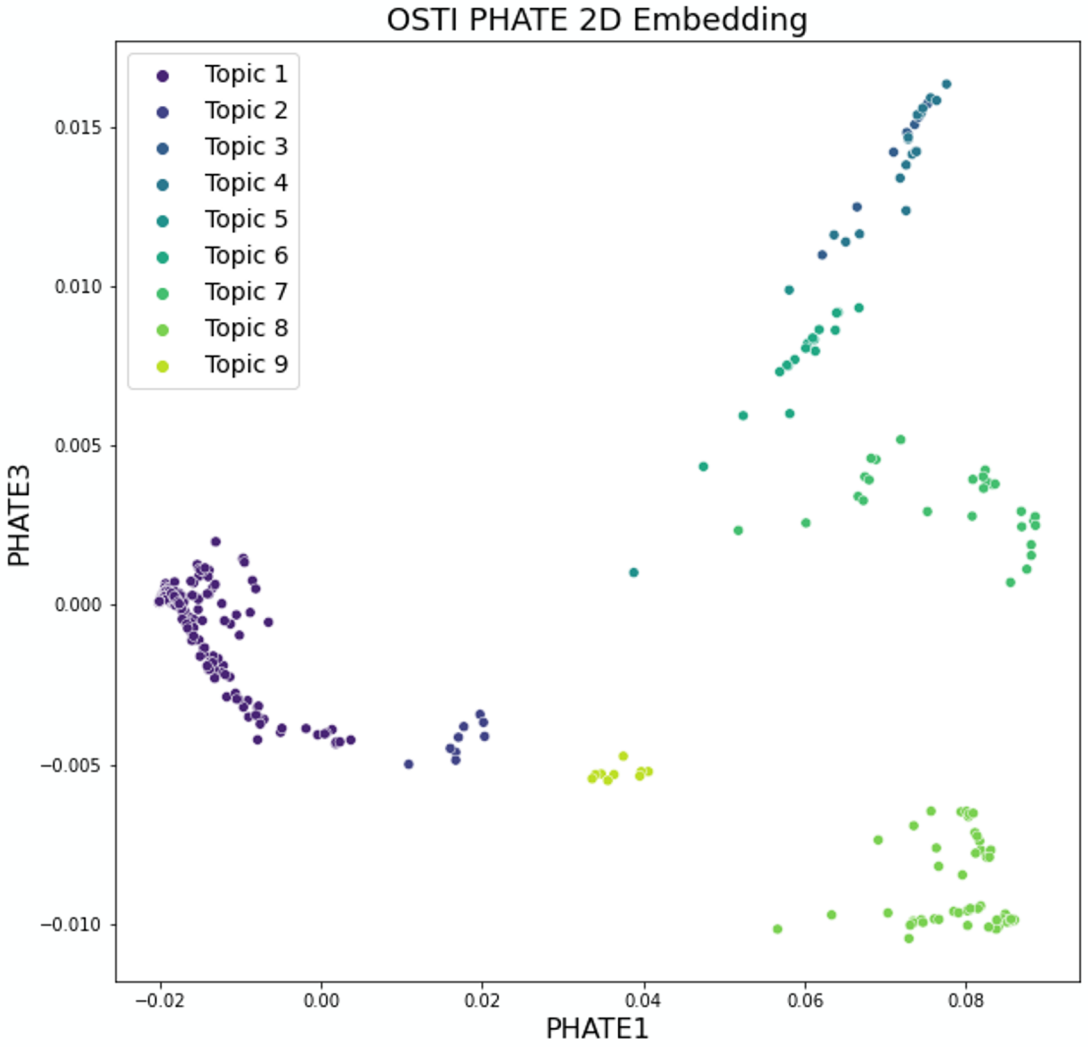




Hellinger Distance

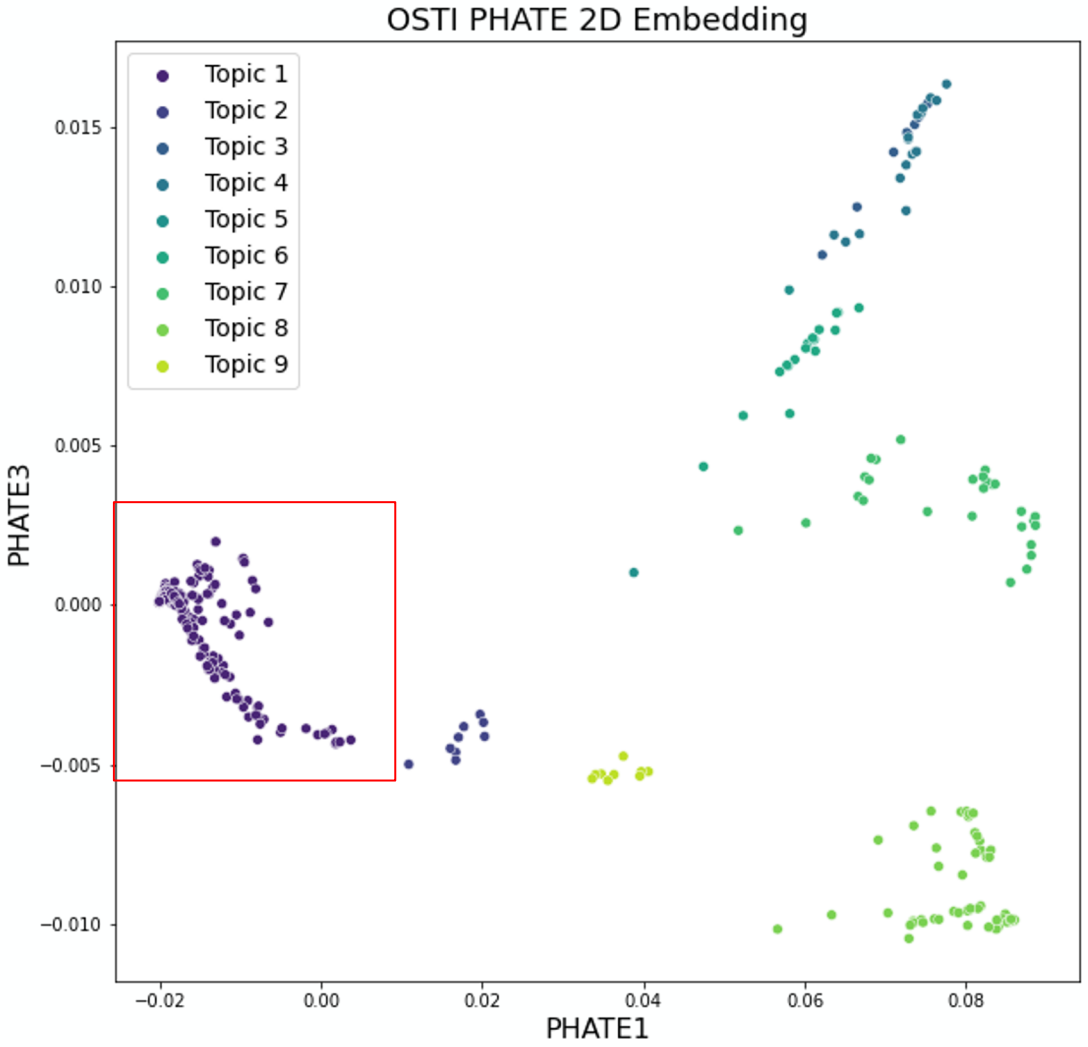
- Hellinger distance generally superior to Euclidean distance for distributions
- Each topic is a distribution on words in the vocabulary
- Using Hellinger metric, kNN graph is used to compute topic trajectories

$$H(p,q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{n=1}^{N} (\sqrt{p_n} - \sqrt{q_n})^2}\,, \qquad 0 \le H(\cdot,\cdot) \le 1$$



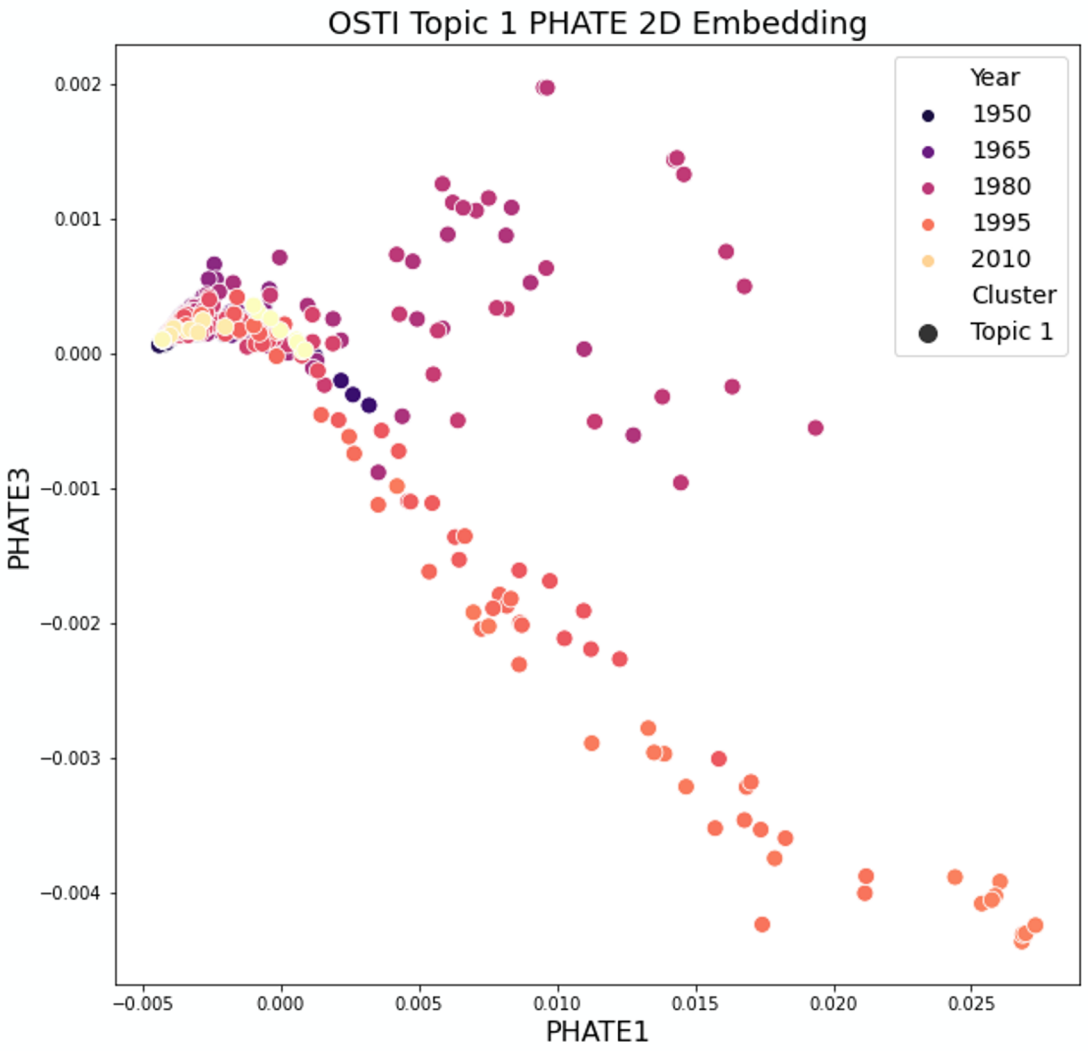
PHATE Embedding

- Low-dimensional embedding for understanding topic evolution
- Captures convergence, divergence, emergence of topics
- Has been shown to have advantages over other low-dimensional embedding methods for temporal data
- Discovered topics are visually separated



PHATE Embedding

- Can use PHATE to investigate specific topics in more detail
- Topic 1 temporal evolution shown over time with varying color
- Clear separation between different decades of research
- Cold war and nuclear treaties may have impacted research trajectories



PHATE Embedding

- Can use PHATE to investigate specific topics in more detail
- Topic 1 temporal evolution shown over time with varying color
- Clear separation between different decades of research
- Cold war and nuclear treaties may have impacted research trajectories

Future Work

- Co-authorship network data
  - Use public co-authorship data to inform dynamic topic learning

- Predictive collaboration
  - Can we predict which authors will collaborate with other authors?
  - Are there anomalous collaborations between authors?

- Information diffusion
  - Dynamic network and topic models combined can be used to model information diffusion in community

References

[1] Wang, Y., Hougen, C., Oselio, B., Dempsey, W., & Hero, A. (2021). A Geometry-Driven Longitudinal Topic Model. *Harvard Data Science Review,* 3(2). https://doi.org/10.1162/99608f92.b447c07e

[2] Blei, D. M., & Lafferty, J.D. (2006). Dynamic topic models. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113-120). Omni Press.

[3] Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, *37* (12), 1482–1492.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Department of Energy / National Nuclear Security Administration under Award Number(s) DE-NA0003921.