Sub-topic and Semantic Sub-structure Extraction via SPLIT: Joint Nonnegative Matrix Factorization (NMF) with Automatic Model Selection

*Maksim E. Eren, ◆Nicholas Solovyev, ◆Ryan Barron, ◆Manish Bhattarai, ◆Ismael D. Boureima, ◆Kim O. Rasmussen, ◆Boian S. Alexandrov
*Analytical Division, ◆Theoretical Division

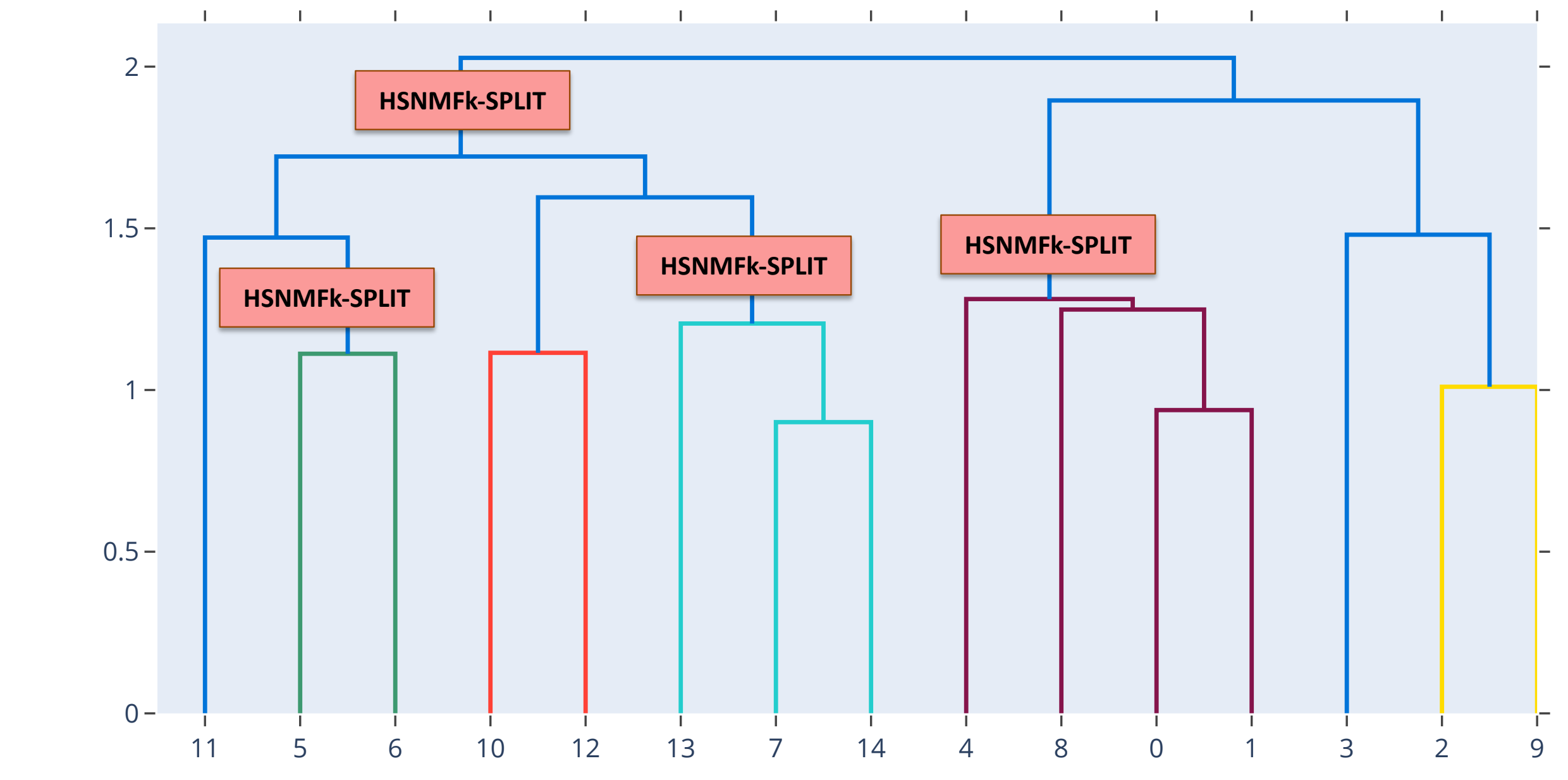
Contact: maksim@lanl.gov

Objective

- Topic modeling is one of the key analytic techniques for organizing and analysis large text corpora.
- We have previously introduced Semantic NMFk[1]: which incorporate the **semantic structure of the text** with the ability to **estimate the number of topics**[2].
- Here, we introduce a new **method for large-scale data analysis**.
- We decompose large **text-document matrix fast using chunks/parts of it and joint factorization**.
- **We hierarchically apply SeNMFk** to extract complex structure of **sub-topics beyond the main themes**.
- We identify corresponding sub-semantic structures that can serve as specific vocabularies – scientific-jargon for local **Name Entities Recognition (NER).**
- We enhance semantic clustering of each topic by **jointly factorizing the arXiv-category - word matrix**.

HSNMFk-SPLIT: Topic and Sub-Topic Modeling Method Designed for Large Corpora, with Hierarchical Application of Semantic NMF with Determination of the Number of Topics

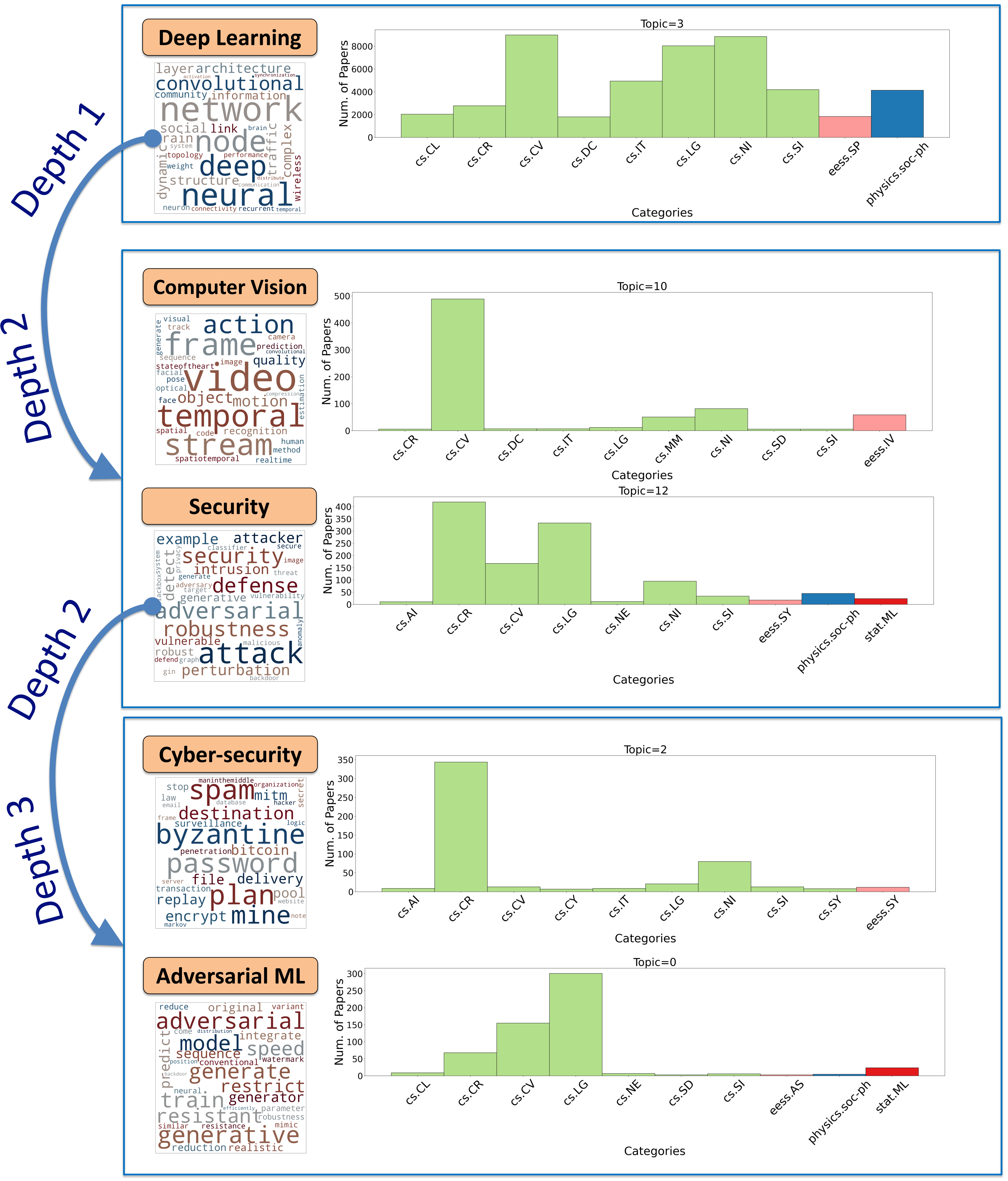
Illustration of Hierarchically Applying our Method



REFERENCES

[1] Maksim E. Eren, Nick Solovyev, Manish Bhattarai, Kim Rasmussen, Charles Nicholas, and Boian S. Alexandrov. 2022. SeNMFk-SPLIT: Large Corpora Topic Modeling by Semantic Non-negative Matrix Factorization with Automatic Model Selection. In ACM Symposium on Document Engineering 2022 (DocEng '22), September 20-23, 2022, San Jose, CA, USA. ACM, New York, NY, USA, 4 pages.

[2] Boian Alexandrov, Velimir Vesselinov, and Kim Orskov Rasmussen. SmartTensors unsupervised ai platform for big-data analytics. Technical report, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2021. LA-UR-21-25064.

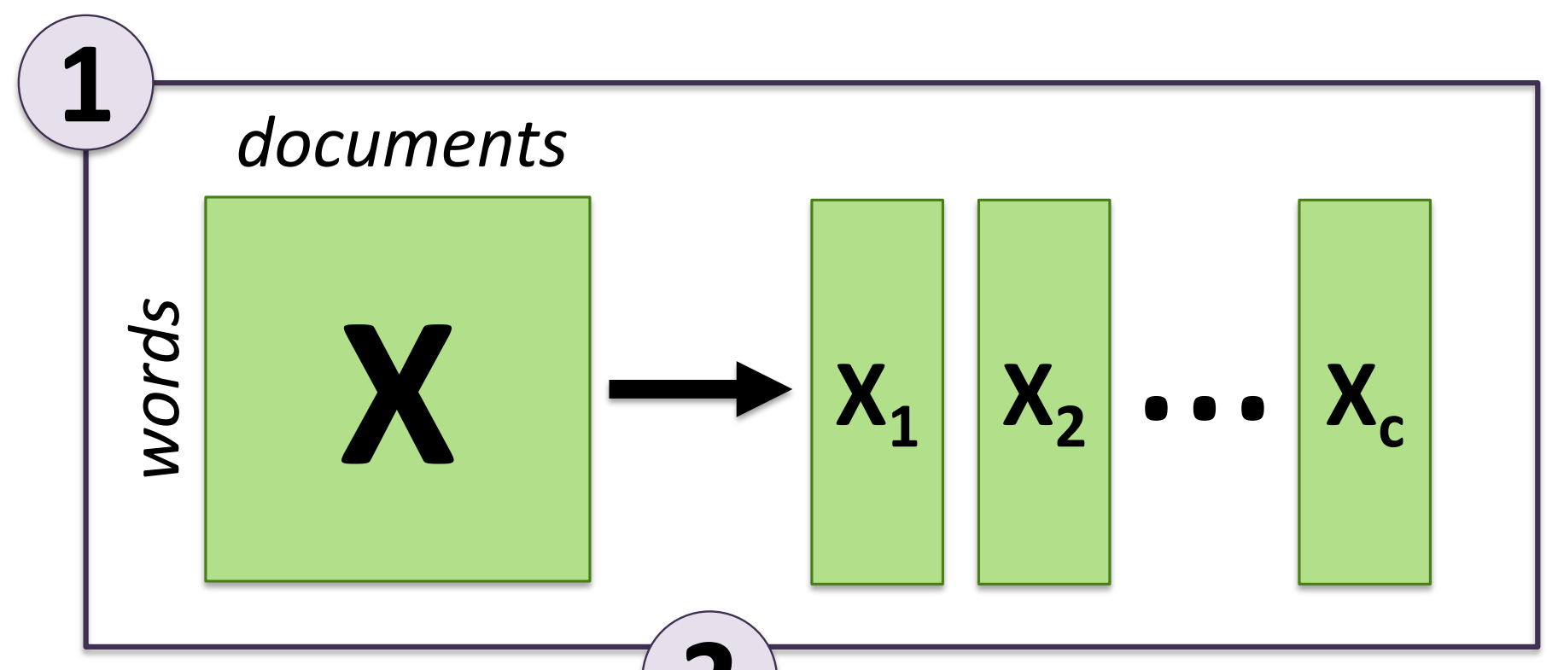


Experiments

- Demonstrate our method by **performing topic modeling on all ~2 million+ papers** posted on **arXiv**.
- Showing the top 10 **arXiv** categories of the papers in each topic. For example:
- Depth 1: Topic 3 describes deep learning methods.
- Depth 2 includes the sub-topics computer vision (topic 10) and security (topic 12).
- Depth 3: includes cyber-security (topic 2) and adversarial ML and robustness (topic 0) in categories cyber-security, computer vision, and language models.

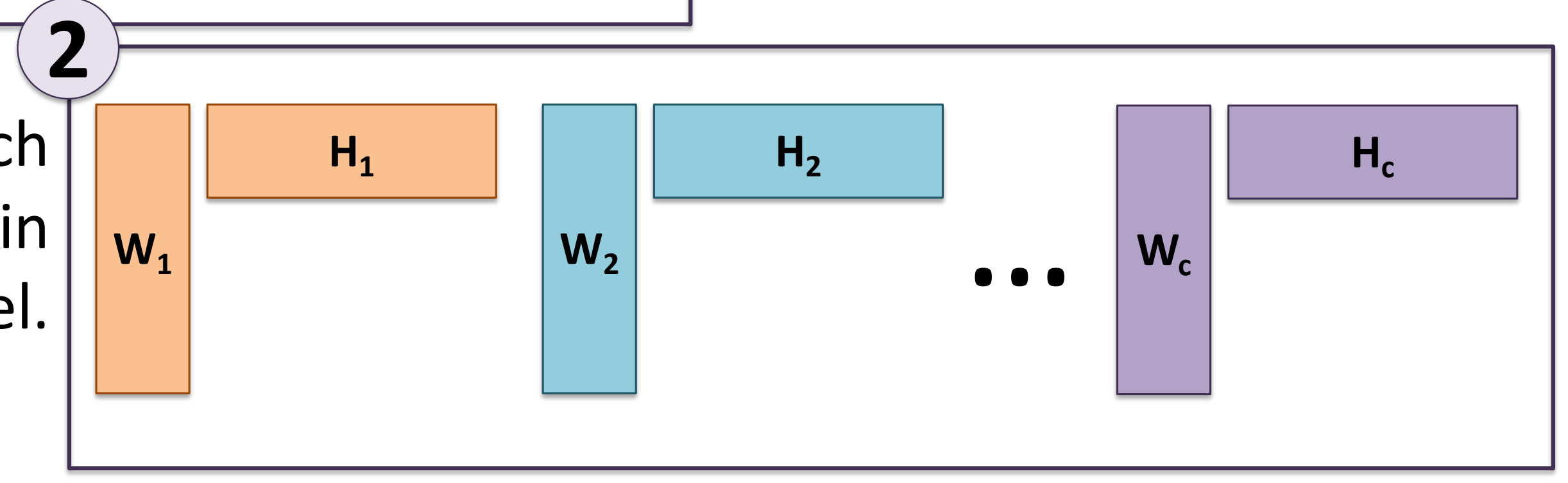
Method

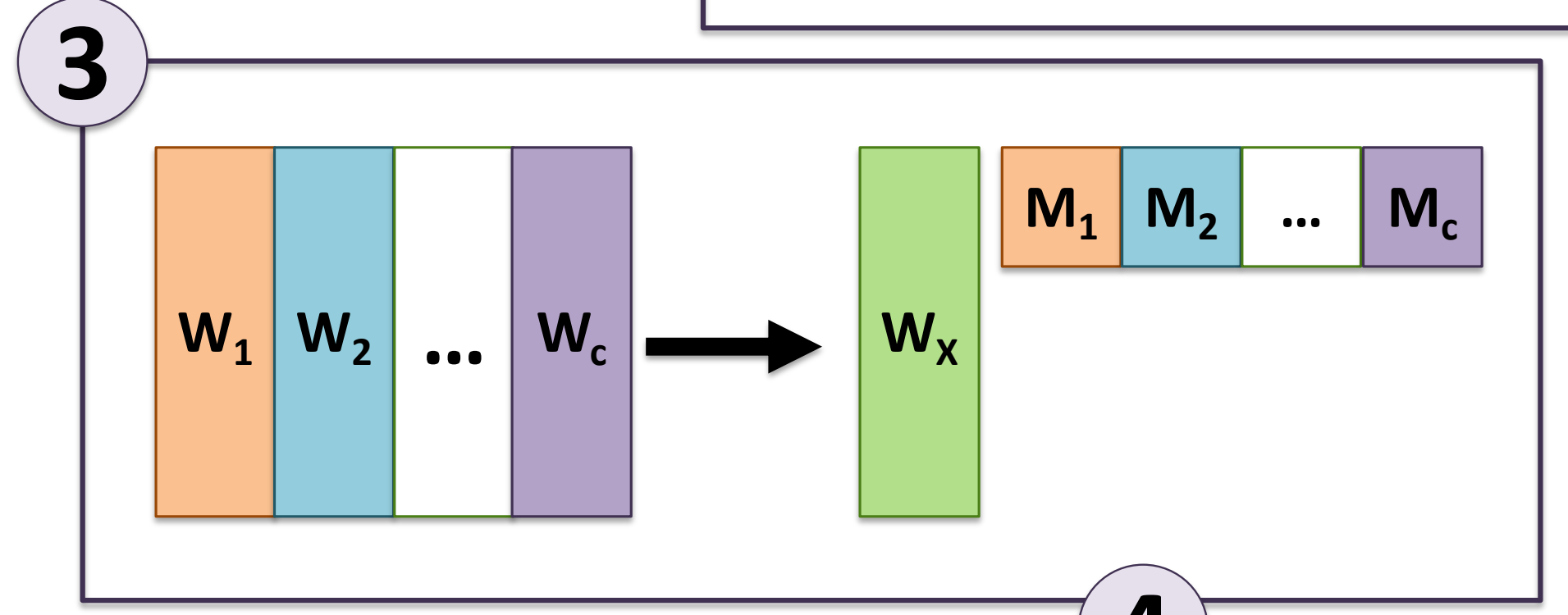
Factorizing Large Matrices via SPLIT



- Split the text-document TF-IDF matrix into C chunks.
- Allows handling smaller data at a time.

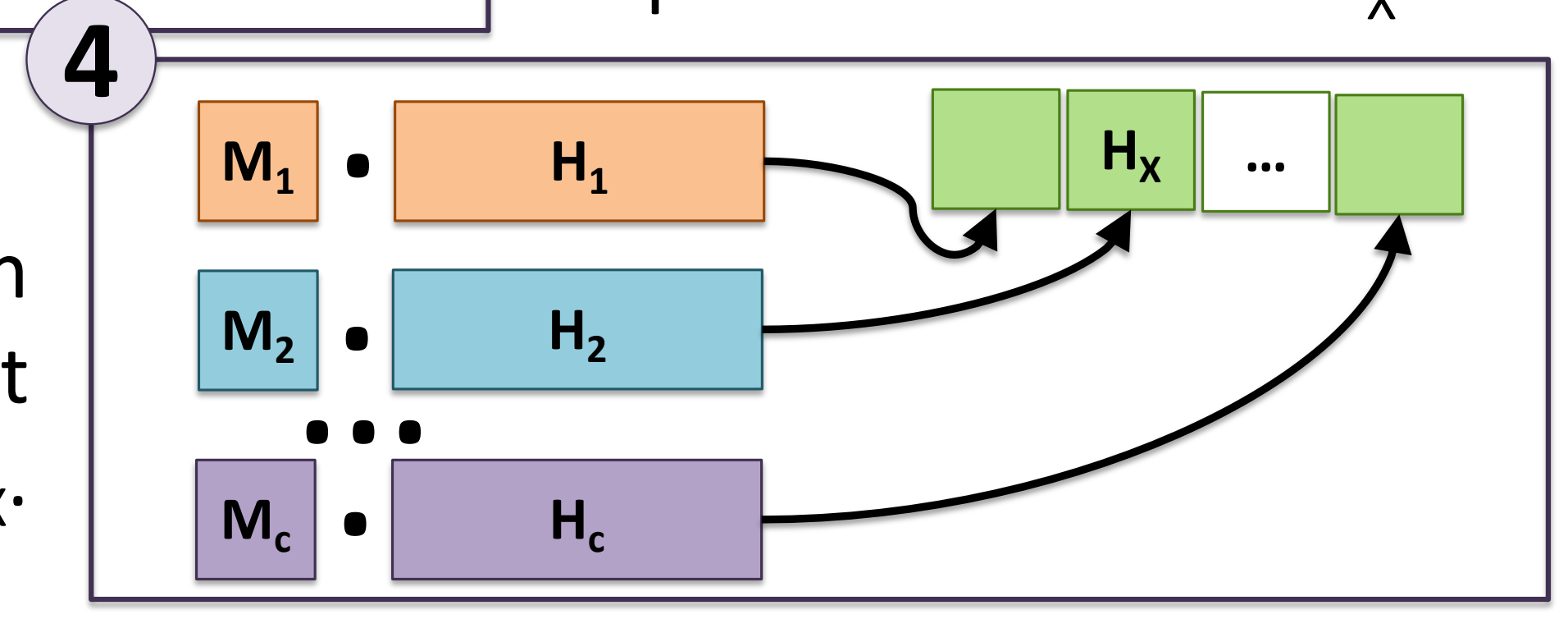
- Factorize each chunk in parallel.



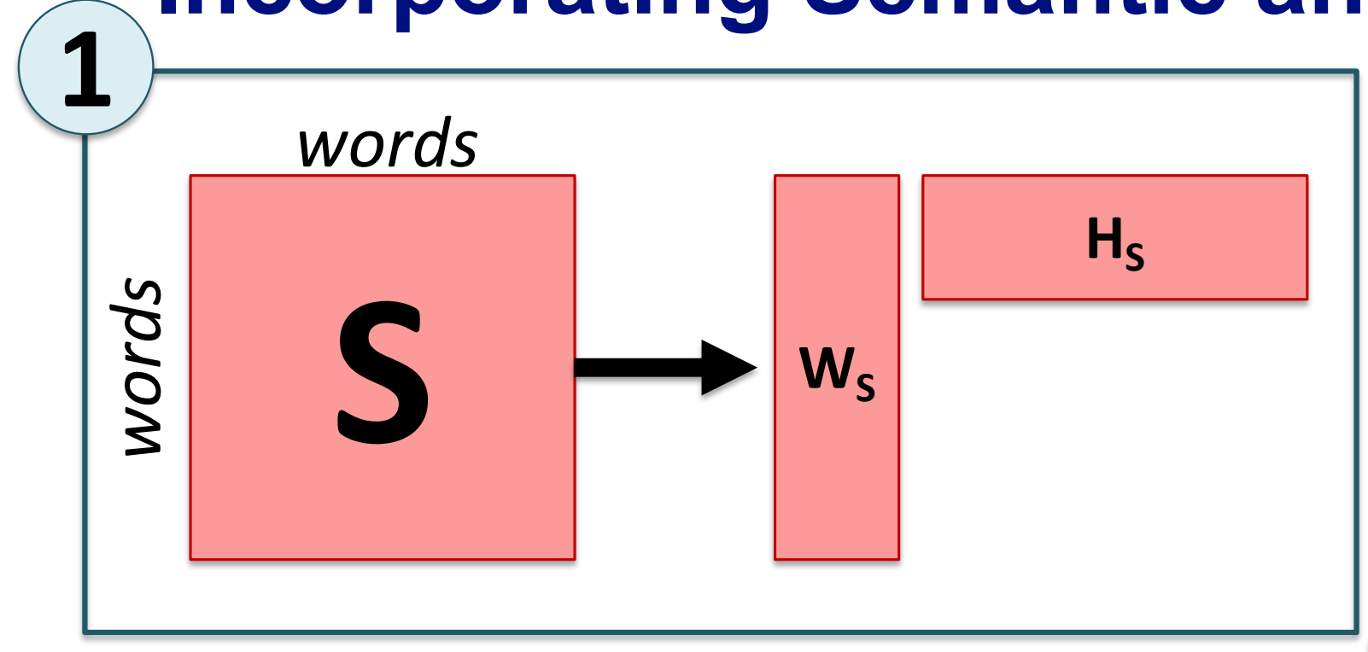


- Concatenate **W**s from each chunk and jointly factorize.
- Obtain the common patterns and final $\mathbf{W}_X$.

- Transfer the common patterns to the document assignments for final $\mathbf{H}_X$.

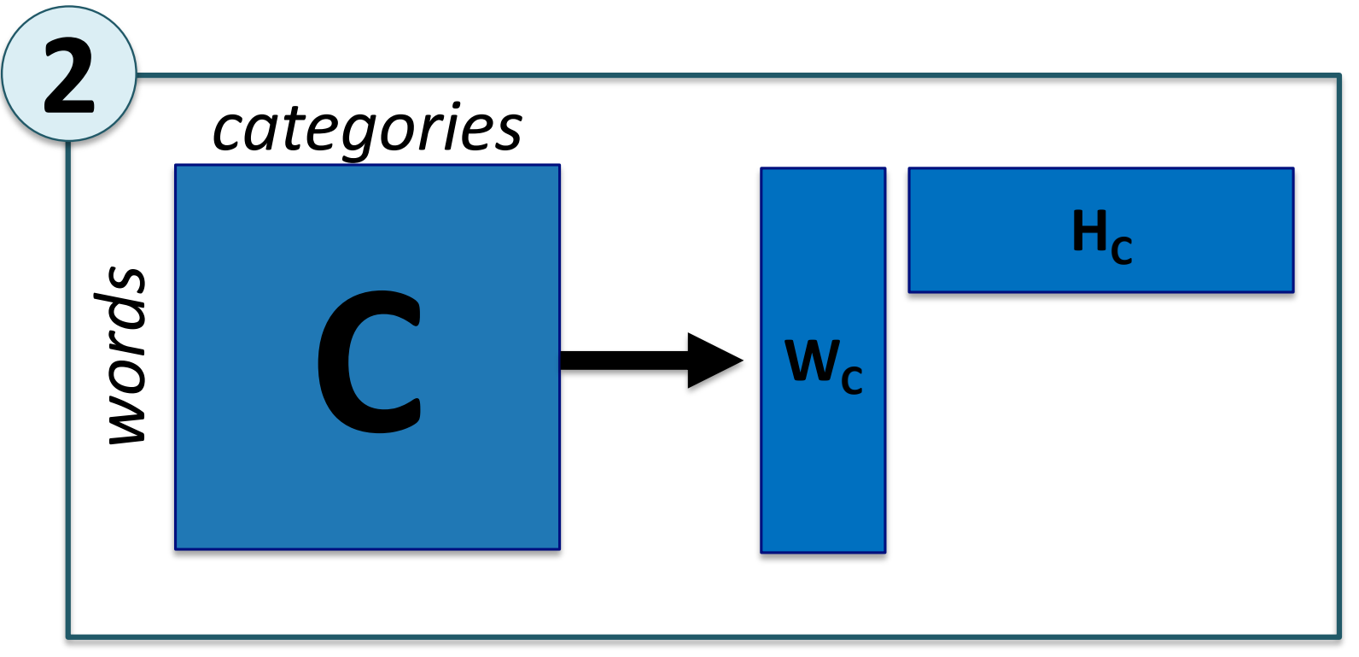


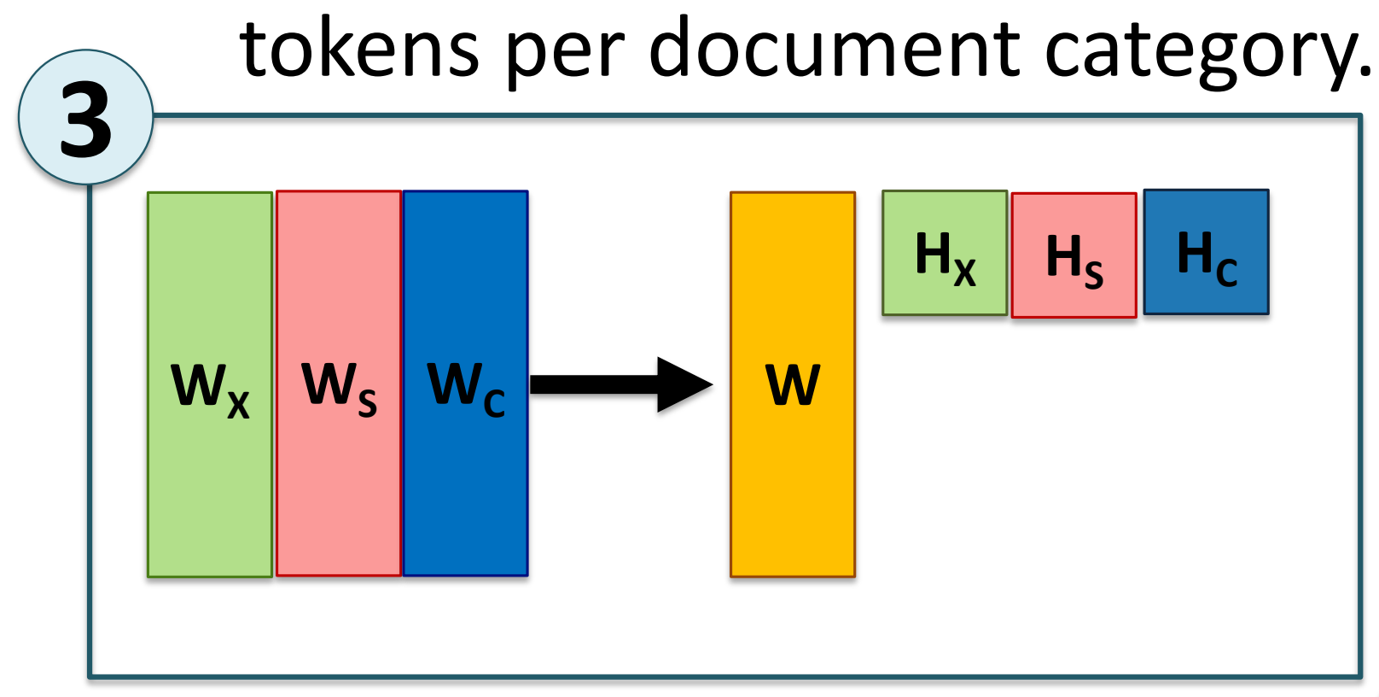
Incorporating Semantic and arXiv-Category Structure



- Factorize the word-context matrix which represents the number of times two words co-occur in a pre-determined window of text.

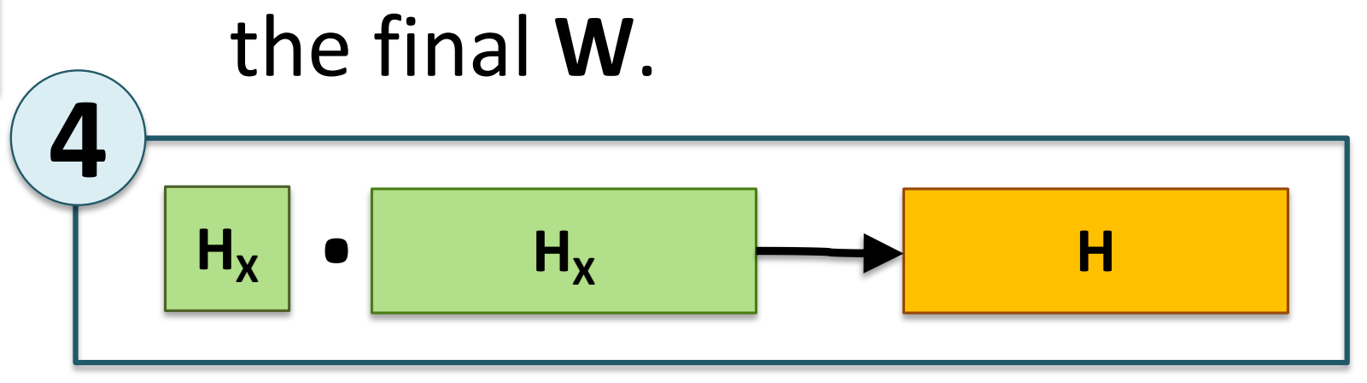
- Factorize the word-arxiv-category matrix, values of which represents the TF-IDF of tokens per document category.





- Jointly factorize the semantic and category patterns to obtain the category identification, and the final **W**.

- Transfer common word patterns to document assignments for final **H**.



ACKNOWLEDGEMENT

This research was funded by DOE National Nuclear Security Administration (NNSA) - Office of Defense Nuclear Nonproliferation R&D (NA-22)

Presented at the *Conference on Data Analysis* (CoDA), Santa Fe, New Mexico. March 7-9, 2023.