

# SLIC: Scientific Leadership Identification and Characterization: Interactive Distillation of Large Single-Topic Corpora of Scientific Papers

Nicholas Solovyev (Theoretical Division, LANL)

Ryan Barron (Theoretical Division, LANL)

**Maksim E. Eren (Advanced Research in Cyber Systems, LANL)**

Kim O. Rasmussen (Theoretical Division, LANL)

Manish Bhattarai (Theoretical Division, LANL)

Ismael D. Boureima (Theoretical Division, LANL)

Boian S. Alexandrov (Theoretical Division, LANL)

October 24-26 2023

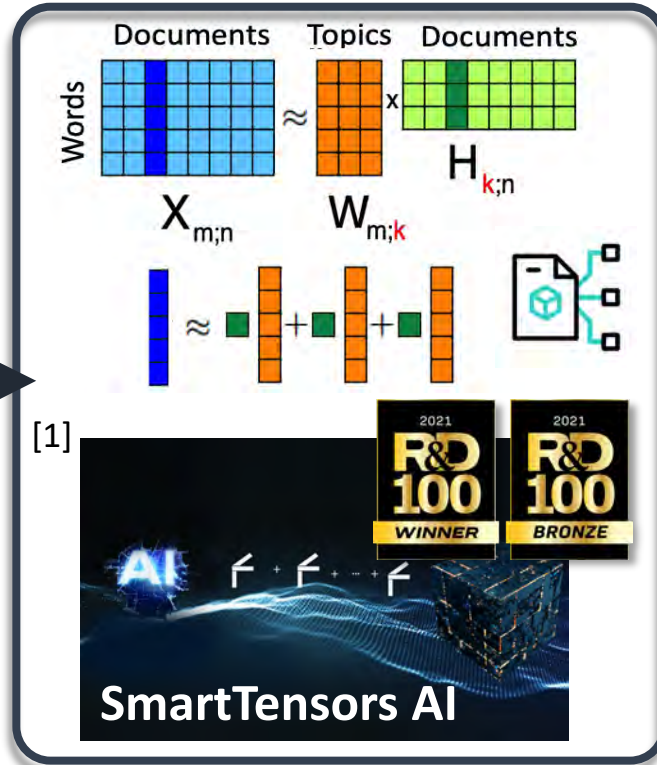
LA-UR-23-30223

# SLIC: Scientific Leadership Identification and Characterization

Non-negative Matrix/Tensor Factorization + Semantics + Determination of the Number of Topics

## Pre-processing

- Document collection
- Remove ambiguities
- Clean documents: **NLP**
- Build tensors



## Post-processing

- Document Clustering
- Topic Evolution
- Paper Ranking
- Author Ranking
- Community Ranking
- Web Hunting
- Report Generation



Extraction of Latent Features from **Big-Data**  
[smart-tensors.LANL.gov](http://smart-tensors.LANL.gov)

# What SLIC can do?

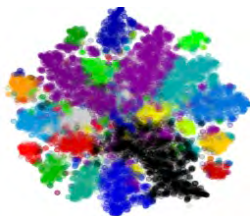


**\*HPC Scale**

**Sub-Topics Extractions:** Word clouds

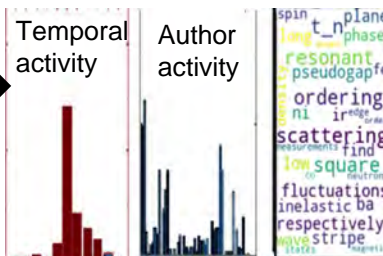


**Document Clustering:** Documents corresponding to different topics

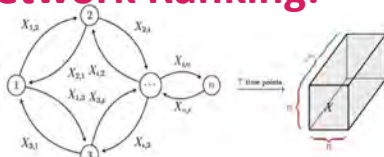


**Temporal activity (Topic Evolution):**

The discovery of iron-based Superconductors in public literature

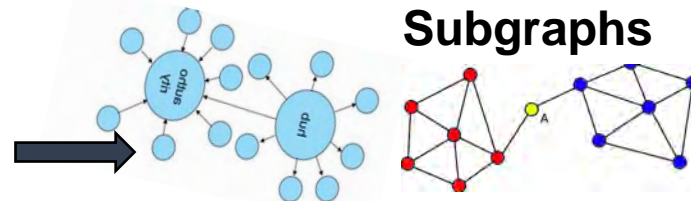


**Social Network Ranking:**



**Co-citation network**

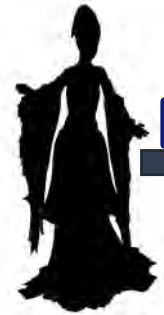
**Co-citation Subgraphs**



**Tensor Embedding**



# Large Software Ecosystem: SLIC ZOO



The Task



SLIC-NER Vocabulary

SME



**2. Cheetah**  
Fast search by keywords



**3. Penguin & iPenguin**  
Matching to S2 IDs



**4. Orca**  
Unique Author IDs



**1. Crocodile**  
Downloading and Reformatting data



Tensor ELF



**5. Vulture**  
NLP cleaning



**10. Bunny**  
Hops: Web hunting



**6. Beaver**  
Building matrices and tensors



**9. Peacock**  
Reports & Figures



**8. Wolf**  
Graph Ranking



**7. Doxie**  
Pick a sub-topic



## Where are we?



Scopus<sup>®</sup>

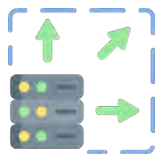
Analyzed 1M papers of Scopus data



Factorized and analyze the whole arXiv: ~ 2M papers



Can run T-ELF on Amazon Cloud



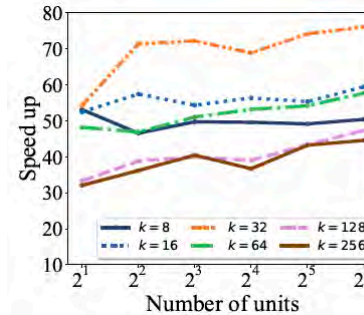
Can run exascale RESCAL and NMFk

# HPC: NMFk & RESCALk

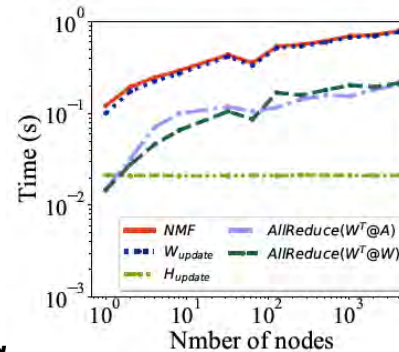
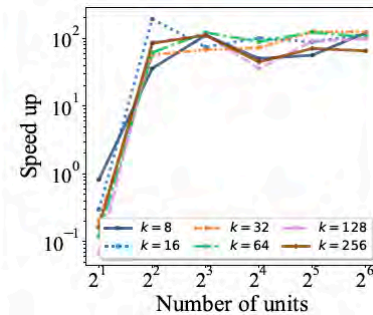
First NMFk and RESCALk distributed code on CPU/GPUs

- Novel distributed algorithm with **out-of-memory support** for NMF for **sparse and dense matrices** operating **across CPU-GPU hardware**.
- **First NCCL communicator** accelerated NMF decomposition tool in distributed GPUs.
- Demonstrate scalability over a **record-breaking**:
  - 340 Terabytes (TB) dense
  - 11 Exabytes (EB) sparse synthetic datasets.
  - **WORLD RECORD: 25k GPUs**
- *Knowledge graph embedding for AI reasoning*

Calculations



Communications



Efficient scaling



22<sup>nd</sup> International Conference on Machine Learning and Applications



December 15-17, 2023

AMLA

Hyatt Regency Jacksonville Riverfront, Florida, USA

## *Interactive Distillation of Large Single-Topic Corpora of Scientific Papers*

The Journal of Supercomputing  
Springer Nature

### *Distributed Out-of-Memory NMF on CPU/GPU Architectures<sup>[2]</sup>*



### *Distributed Out-of-Memory SVD on CPU/GPU Architectures<sup>[4]</sup>*

21<sup>st</sup> IEEE International Conference on Machine Learning and Applications



December 12-14, 2022

AMLA

Atlantis Hotel, Bahamas

### *One-Shot Federated Group Collaborative Filtering<sup>[6]</sup>*



### *Distributed non-negative RESCAL with Automatic Model Selection for Exascale Data<sup>[3]</sup>*



March 7 - 9, 2023 | Santa Fe, New Mexico  
Exploring Data-Focused Research across the Department of Energy

### *Sub-topic and Semantic Sub-structure Extraction via SPLIT: Joint Nonnegative Matrix Factorization (NMF) with Automatic Model Selection<sup>[5]</sup>*

The 21<sup>st</sup> ACM Symposium on Document Engineering

DocEng

24-27 August 2021  
Limerick, Ireland



### *SeNMFk-SPLIT: Large Corpora Topic Modeling by Semantic Non-negative Matrix Factorization with Automatic Model Selection<sup>[7]</sup>*



# Objective



Highly specific datasets of scientific literature are important for both research and education, as well as for training large language models.

---



Creating a such dataset with hand is time consuming and prone to errors.

---



New tool based on machine learning (ML) for **constructively generating large-scale specific/targeted datasets of scientific literature.**

---



Novel framework based on combination of information retrieval, word embeddings and large language model, and accurate topic modelling.

---

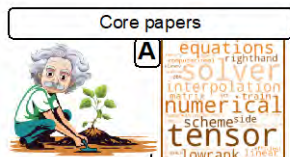


**Human-in-the-loop** assisted ML: interactive, user-driven approach, we empower users to steer the topic extraction process directly, ensuring the results are tailored to their specific requirements

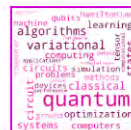


# Method Summary

(1) Subject-matter-expert selects handful of papers



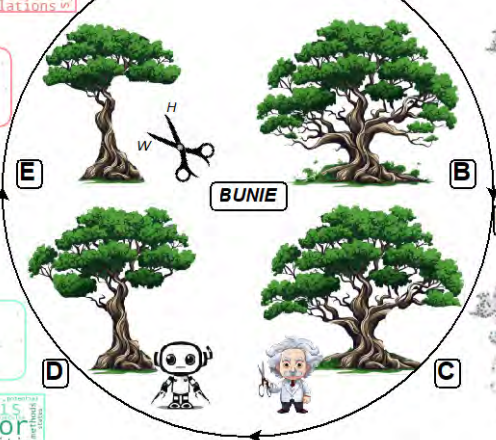
(2) Citation graph used to expand the dataset



(5) Topic modeling based cleaning



Hop Expansion



Human-in-the-loop Pruning



(4) Distance over the hyperspace based cleaning



(3) Select relevant papers by hand



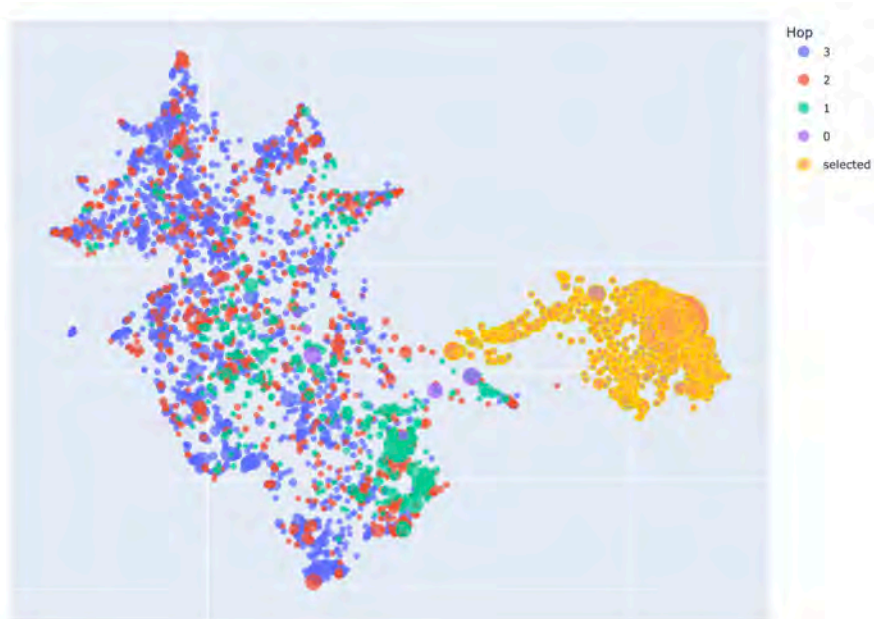
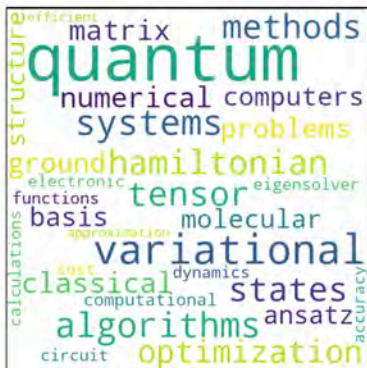
# GUI Demo

**Data Controls**

Plot Labels: Hop  X

Highlight Core:

Delete Selected Papers



<input type="checkbox"/>	selected	id	eid	dol	title	abstract	year	authors	author_ids	affil
<input type="checkbox"/>	false	0	2-s2.0-85137156346	10.1016/j.jcp.2022.111562	A parallel low-rank solver for ... Continuum Vlasov simulation...		2022	Allmann-Rahn F.;Grauer R.;...	57209974131;7003609086;2...	{'600
<input type="checkbox"/>	false	1	2-s2.0-85102687107	10.1137/19m1281435	Regularization of poisson-bol...	In this paper, we present a n...	2021	BENNER P.;KHOROMSKAIA...	22733643800;17135342400;...	{'600
<input type="checkbox"/>	false	2	2-s2.0-85061235760	10.1007/s00162-019-00485-z	On identification of self-simil...	A study on the application of ...	2019	von Larcher T.;Klein R.	11440638600;7404358451	{'122
<input type="checkbox"/>	false	3	2-s2.0-84869825516	10.1137/110833142	Solution of linear systems an...	Tensors arise naturally in hig...	2012	Oseledets I.V.;Dolgov S.V.	8529104000;57200590946	{'601

# Thank you!

## Questions?

Contact: [maksim@lanl.gov](mailto:maksim@lanl.gov)

[maksimeren.com](http://maksimeren.com)

Paper Accepted to the  
**2023 IEEE ICMLA**

### Interactive Distillation of Large Single-Topic Corpora of Scientific Papers

Nick Solovyyev  
*Theoretical Division, LANL*  
Los Alamos, USA  
nks@lanl.gov

Maksim E. Eren  
*Advanced Research in Cyber Systems, LANL*  
Los Alamos, USA  
maksim@lanl.gov

Ryan Barron  
*Theoretical Division, LANL*  
Los Alamos, USA  
barron@lanl.gov

Kim Ø. Rasmussen  
*Theoretical Division, LANL*  
Los Alamos, USA  
kor@lanl.gov

Manish Bhattarai  
*Theoretical Division, LANL*  
Los Alamos, USA  
mcodeps@ccimn@lanl.gov

Boian S. Alexandrov  
*Theoretical Division, LANL*  
Los Alamos, USA  
boian@lanl.gov

**Abstract**—Highly specific datasets of scientific literature are important for both research and education. However, it is difficult to build such datasets at scale. A common approach is to build these datasets reflectively by applying topic modeling on an established corpus and selecting specific topics. A more robust but time-consuming approach is to build the dataset constructively in which a subject matter expert (SME) handpicks documents. This method does not scale and is prone to error as the dataset grows. Here we showcase a new tool, based on machine learning, for constructively generating targeted datasets of scientific literature. Given a small initial “core” corpus of papers, we build a citation network of documents. At each step of the citation network, we generate text embeddings using the Transformer-generated science-specific large language model SciNCL [Ostendorff, Mulke, et al. “Neighbourhood contrastive learning for scientific document representations with citation embeddings.” arXiv preprint arXiv:2205.06671 (2022).] and visualize the embeddings through dimensionality reduction. Papers are kept in the dataset if they are “similar” to the core or are otherwise novelty pruned through human-in-the-loop selection. Additional insight into the papers is gained through sub-topic modeling using SciNMF. We demonstrate our new tool for literature review by applying it to two different fields in machine learning.

**Index Terms**—transformers, nlp, non-negative matrix factorization, data visualization

streamlines the literature review task with a user-friendly and intuitive system while enhancing the specificity of the papers of interest using ML techniques and integrated human-in-the-loop procedures.

In this work, we contribute a novel approach to the scientific dataset expansion problem by jointly integrating Transformer-based document text embeddings with human-in-the-loop pruning to generate targeted scientific datasets. We then use non-negative matrix factorization (NMF) with automatic model determination (NMF) for modeling the topics in these papers to further refine our datasets [1]. Our approach is unique in its inclusion of a human-in-the-loop for enhancing and distilling the extracted topics, such that the corpus of papers is narrowed down via an interactive process. To the best of our knowledge, this iterative method is the first of its kind to offer users the ability to analyze the topic modeling results and apply their feedback to enhance the literature review procedure by steering the ML output. The feedback loop enables the users to grow and refine the results until a targeted dataset of a specific size is reached, providing a unique and interactive solution to large-scale literature review.

# References

- [1] 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.
- [2] Boureima, I., Bhattarai, M., Eren, M., Skau, E., Romero, P., Eidenbenz, S., & Alexandrov, B. (2022). Distributed Out-of-Memory NMF of Dense and Sparse Data on CPU/GPU Architectures with Automatic Model Selection for Exascale Data. Springer Nature. The Journal of Supercomputing.
- [3] Manish Bhattarai, Namita kharat, Ismael Boureima, Erik Skau, Benjamin Nebgen, Hristo Djidjev, Sanjay Rajopadhye, James P. Smith, Boian Alexandrov, Distributed non-negative RESCAL with automatic model selection for exascale data, Journal of Parallel and Distributed Computing, Volume 179, 2023, 104709, ISSN 0743-7315, <https://doi.org/10.1016/j.jpdc.2023.04.010>.
- [4] Boureima, I., Bhattarai, M., Eren, M. E., Solovyev, N., Djidjev, H., & Alexandrov, B. S. (2022). Distributed Out-of-Memory SVD on CPU/GPU Architectures. IEEE HPEC Conference 2022 with Outstanding Paper Award.
- [5] Eren, M.E., Nicholas, S., Barron, R., Bhattarai, M., Boureima, I.D., Rasmussen, K.O., and Alexandrov, B.. Sub-topic and Semantic Sub-structure Extraction via SPLIT: Joint Nonnegative Matrix Factorization (NMF) with Automatic Model Selection. CoDA '23: Conference on Data Analysis, March 7-9, 2023, Santa Fe, New Mexico, USA.
- [6] M. E. Eren and M. Bhattarai and N. Solovyev and L. E. Richards and R. Yus and C. Nicholas and B. S. Alexandrov, "One-Shot Federated Group Collaborative Filtering," 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 2022, pp. 647-652, doi: 10.1109/ICMLA55696.2022.00107. Awarded Best M.S. Research at 2023 UMBC CSEE Research Day.
- [7] 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.