# AUTOMATIC DATABASE CONSTRUCTION FOR MLSYS PAPERS


## ABSTRACT

The rapid expansion of research in Machine Learning Systems (MLSys) necessitates an organized approach to track its evolution, key contributors, and emerging topics. Traditional literature reviews struggle to keep pace with the increasing volume of publications, motivating the need for automated methods to curate and analyze MLSys research. This work presents a machine learning-driven approach to constructing and studying a curated database of MLSys papers and uncovers insights into research trends within the MLSys community.


## 1 INTRODUCTION

The field of Machine Learning Systems (MLSys) is rapidly evolving, with new research contributions emerging. Traditional literature review methods are unable to keep up with the increasing volume of publications, highlighting the need for *automated approaches to curate and analyze MLSys research*. This work introduces a machine learning-driven methodology to construct and study a curated database of MLSys papers. Our approach incorporates automated data collection and analysis techniques to derive meaningful insights into research trends and author networks.

## 2 AUTOMATIC DATABASE CONSTRUCTION

**Step 1: Seed Papers Selection.** Curated seed papers from MLSys conference proceedings (2018-2024) and the Awesome System for Machine Learning repository.

**Step 2: Database Expansion.** Collected references and citations from seed papers using the Semantic Scholar API and applied automated filtering techniques.

**Step 3: Relevant Papers Filtering.** We first employed well-established NLP models, SPECTER2 and Sent2Vec, to extract document-level and sentence-level embeddings from collected papers. K-Means clustering was applied to seed papers to establish clusters representing core topics. Each cluster's centroid was stored, and the radius was determined by the maximum distance from the centroid to any point within the cluster. New papers from the expanded database were evaluated using KNN classification. A paper was deemed relevant if its embedding was within 80% of the cluster's radius; otherwise, it was marked as irrelevant.

## 3 DATABASE ANALYSIS AND INSIGHTS

We propose to answer the following questions based on our constructed MLSys papers database for a better understanding of the research trends and other meaningful information in this research area.

- **Q1:** What are the most prominent research themes in MLSys, and how have they evolved over time?
- **Q2:** Which authors, institutions, and venues are most influential in MLSys research?
- **Q3:** Can we predict emerging trends and identify promising future research directions?

For **Q1**, prominent research themes identified include *distributed training, hardware acceleration, and model optimization*. Evolution of these themes is traced through shifts in cluster centroids over time. Specifically, distributed training has grown consistently over the years, while hardware acceleration saw a significant rise starting in 2020, coinciding with advancements in specialized AI hardware. Model optimization continues to be a central focus but is now increasingly oriented towards efficiency improvements.

For **Q2**, graph-based citation analysis such as PageRank is applied to identify influential authors, institutions, and venues by examining central nodes in citation graphs. Prominent authors include researchers from *Google, Microsoft Research, and Stanford University*, while the most cited venues are *MLSys, NeurIPS, and ICML*. High-impact publications are generally produced through collaborations between top-tier research institutions.

For **Q3**, predictive analysis was performed by applying time-series models to track the growth of clusters. Results indicate increased research in *system optimization and speedup areas*, including model efficiency, large-scale distributed training, and hardware optimization.

## 4 CONCLUSION

This work demonstrates the potential of machine learning techniques for structuring and analyzing large-scale academic datasets. Our approach enables more effective literature reviews and predictive analysis of future trends. Future directions include enhancing data pipelines, expanding the framework to interdisciplinary research domains, and improving the interpretability of generated insights.