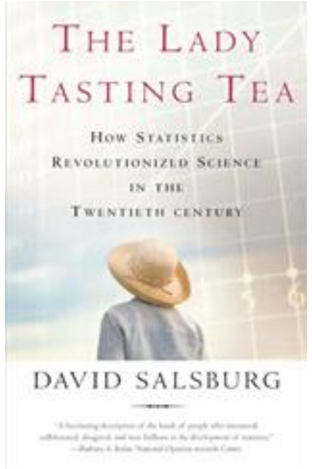
UCLA Undergraduate
RESEARCH WEEK

Research on Research on Research: Analyzing Historical Trends in Statistical and Computational Research from the 1990s to Modern Day

Naren Prakash

Motivation

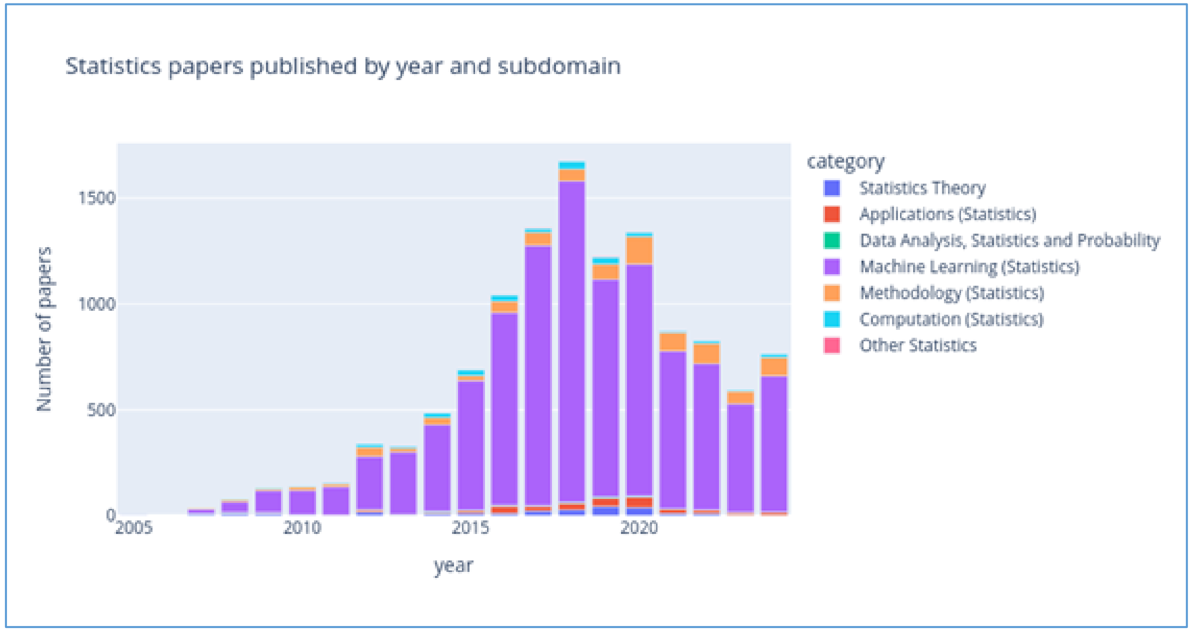
- How has the current research landscape for statistical and computational fields changed in the past 30 years?
- Can we trace the development of ideas in more “recent” fields such as generative AI and reinforcement learning?

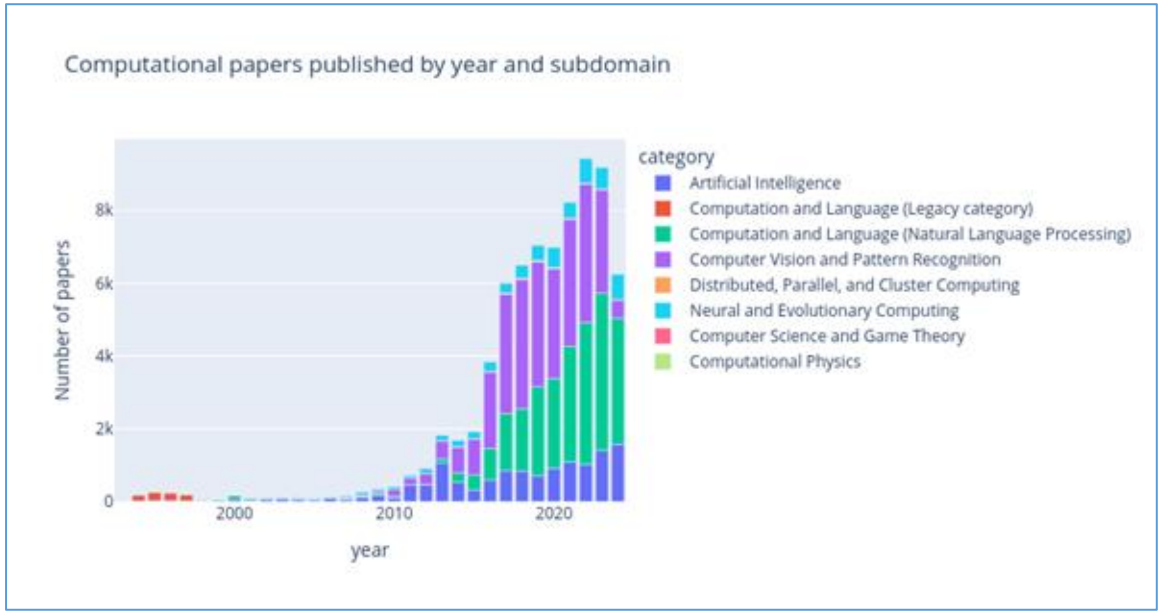

THE LADY TASTING TEA
HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY
DAVID SALSBURG

Dataset

- 136,238 observations and 10 columns
- Column names: id, title, category, category code, published date, updated date, authors, first author, summary, and summary word count
- Originally scraped from arXiv, dataset from Kaggle
- For initial analysis: partitioned into two datasets (one statistical, one computational)

EDA Plots


Statistics papers published by year and subdomain


Computational papers published by year and subdomain

Initial Results: Statistical

- Largest publication increases: Machine Learning, Methodology, Applications
- Top 3 fields remained Machine Learning, Methodology, and Application/Computation

Initial Results: Computational

- Largest publication increases: Artificial Intelligence, Computer Vision, Neural and Evolutionary Computing

- Top 3 fields remained Artificial Intelligence, Computation and Language (NLP), Computer Vision and Pattern Recognition

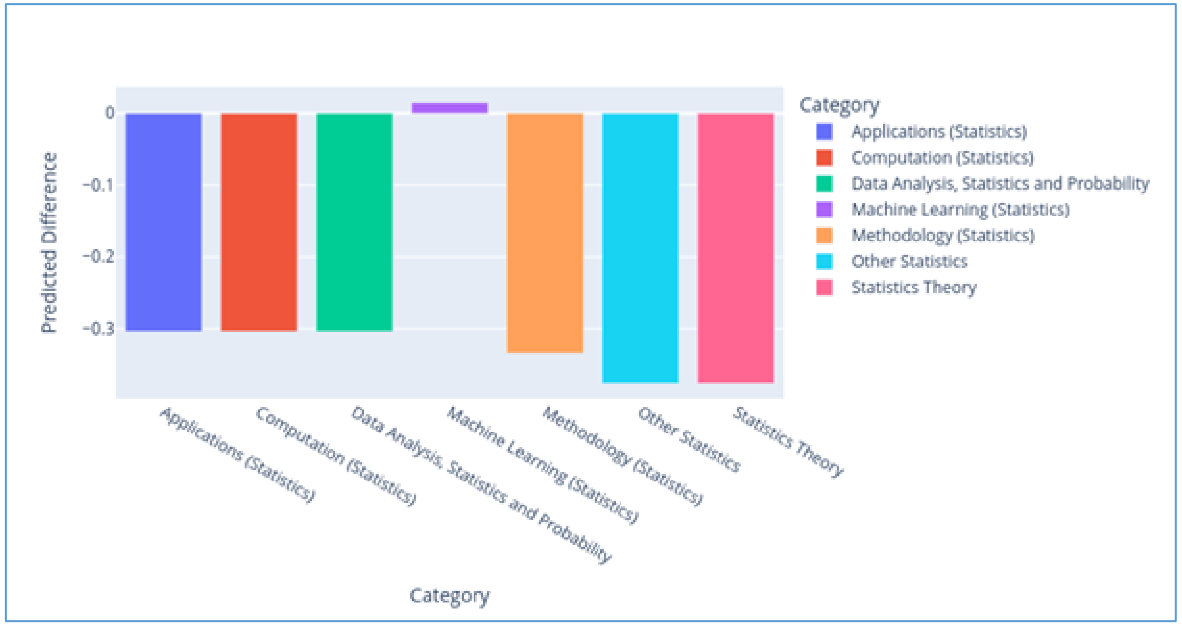
Predictive Model

- Time series model created using a LightGBM regressor
  - Monthly data from 1994 to 2024

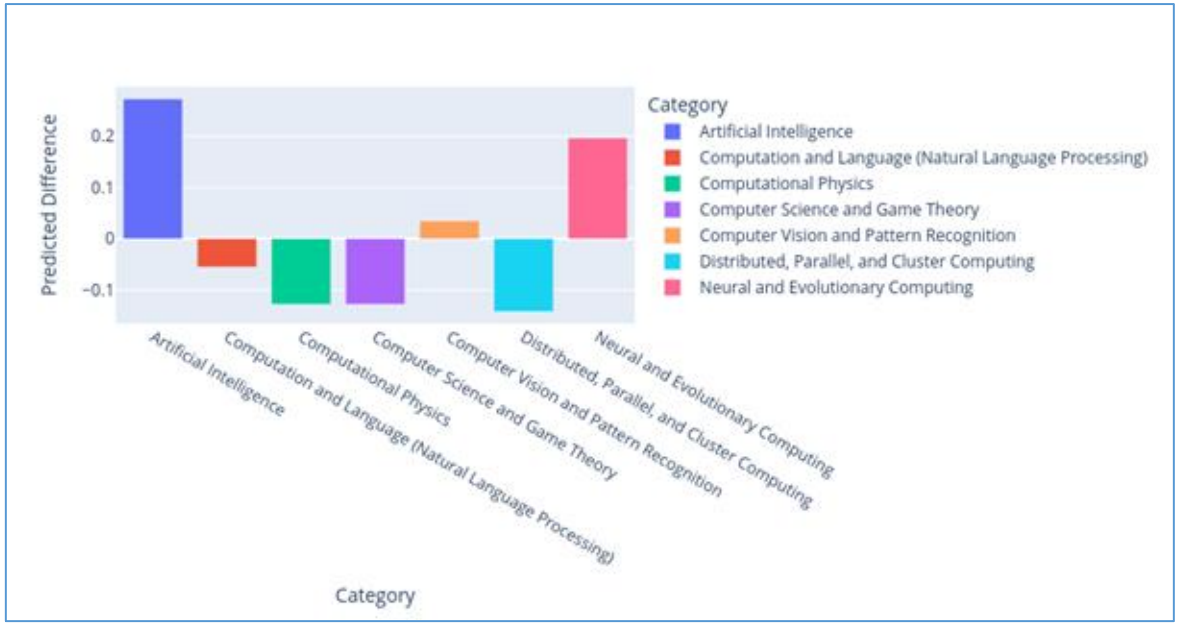
- Aim to predict research output for the first half of 2025

Prediction Results

Statistical



Computational



Digging Deeper: Ideological Progression

- How can we analyze by content and not just category?

- Want to use text mining techniques such as:
    - n-grams
    - topic modeling techniques
    - clustering

BERTopic: Extraction and Labeling

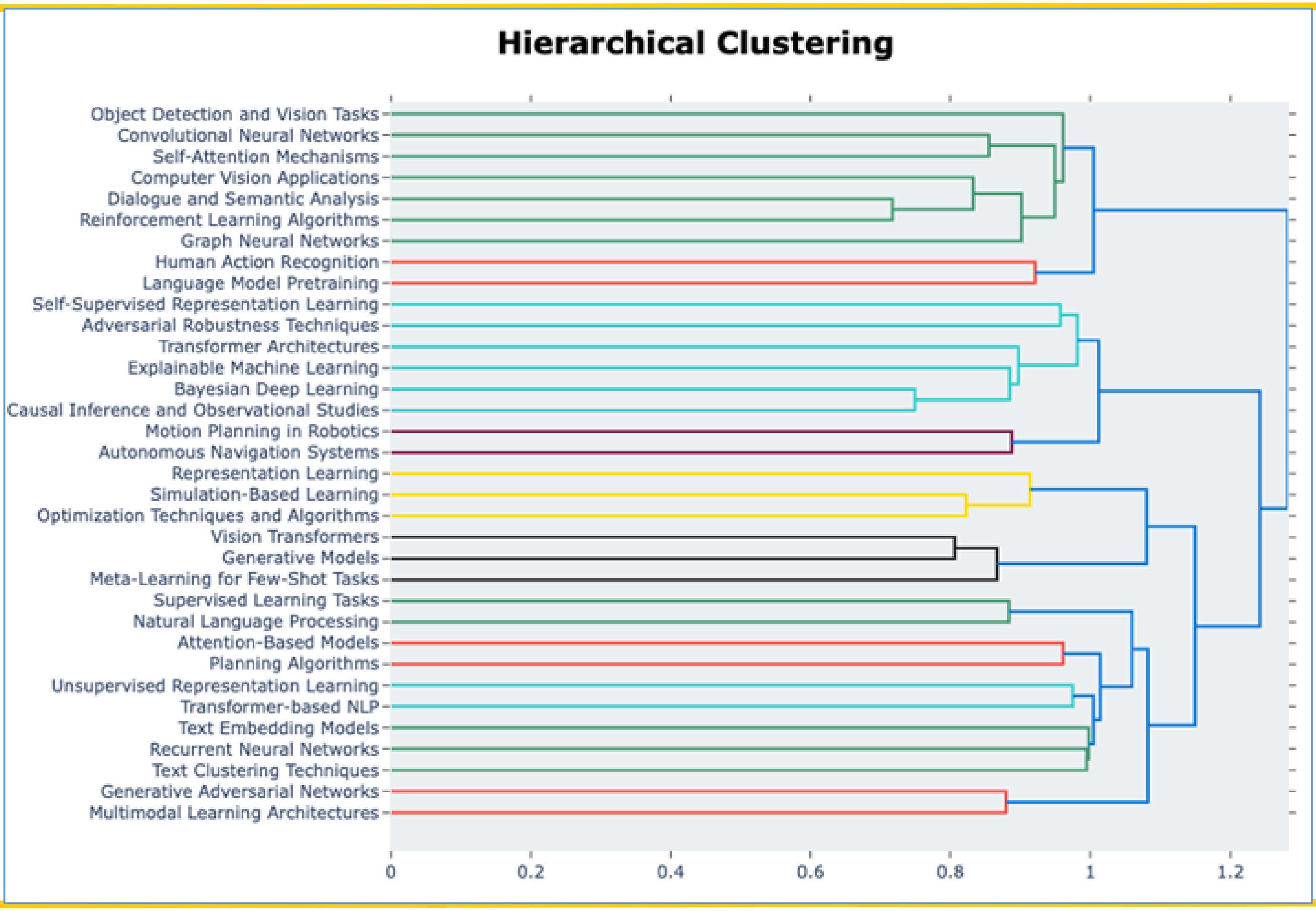
- Used paper abstracts to extract key words and phrases

- Why BERTopic?
  - Contextual understanding from Large Language Models
  - Incorporation of traditional text mining techniques
    - TF-IDF, n-grams, etc.

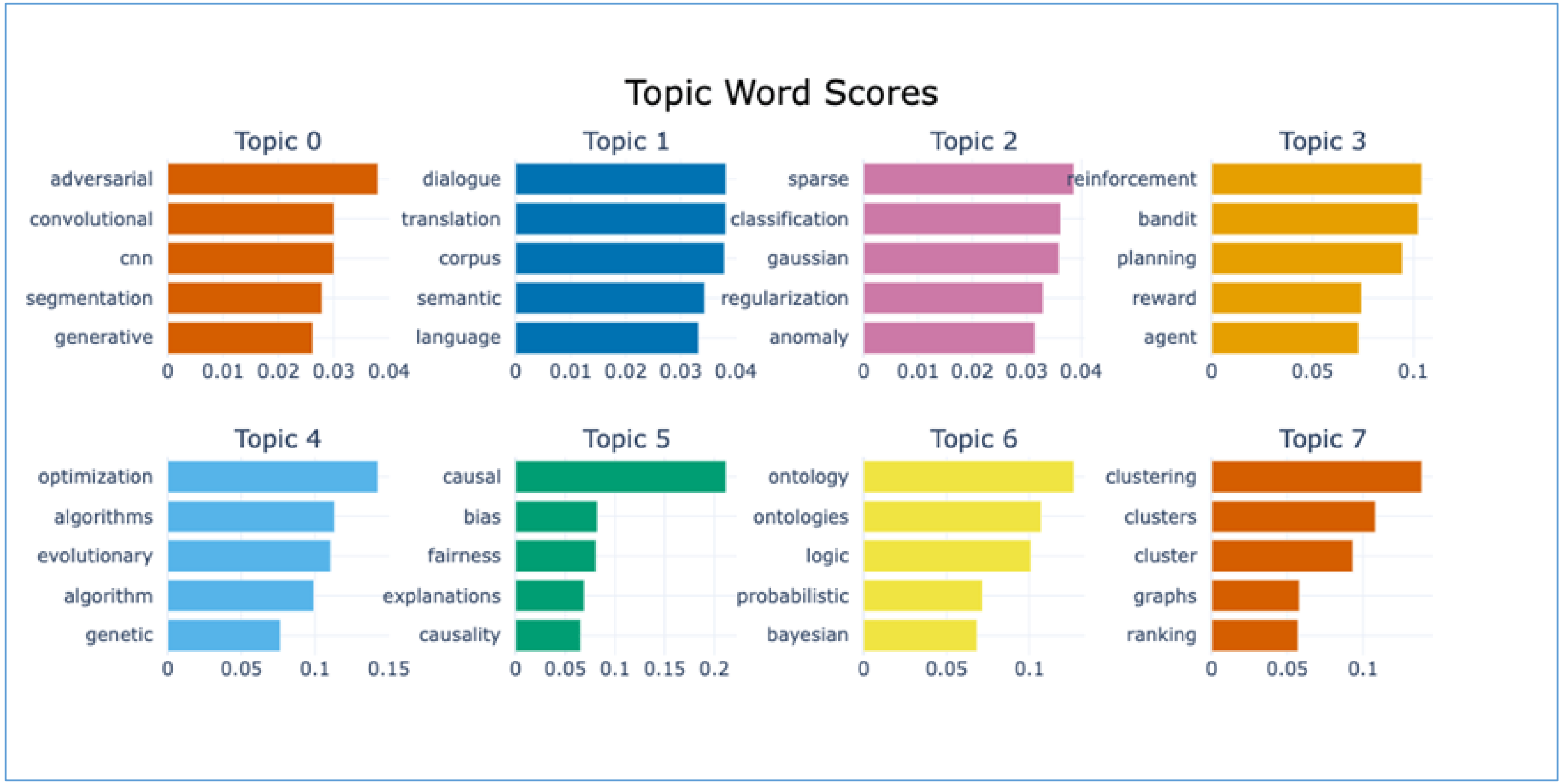
How do we find these labels?

- BERTopic Pipeline:
  - Dimension reduction
    - UMAP (Uniform Manifold Approximation and Projection)
  - Clustering
    - HDBSCAN
  - Topic Modeling
    - CountVectorizer, c-TF-IDF, KeyBERT, ChatGPT

How do we find these labels?



Labels (35 final topics)



0: Adversarial Robustness in Convolutional Networks

1: Dialogue Systems and Machine Translation

2: Sparse Modeling and Gaussian Classification

3: Reinforcement Learning and Bandit Algorithms

4: Optimization Algorithms in Machine Learning

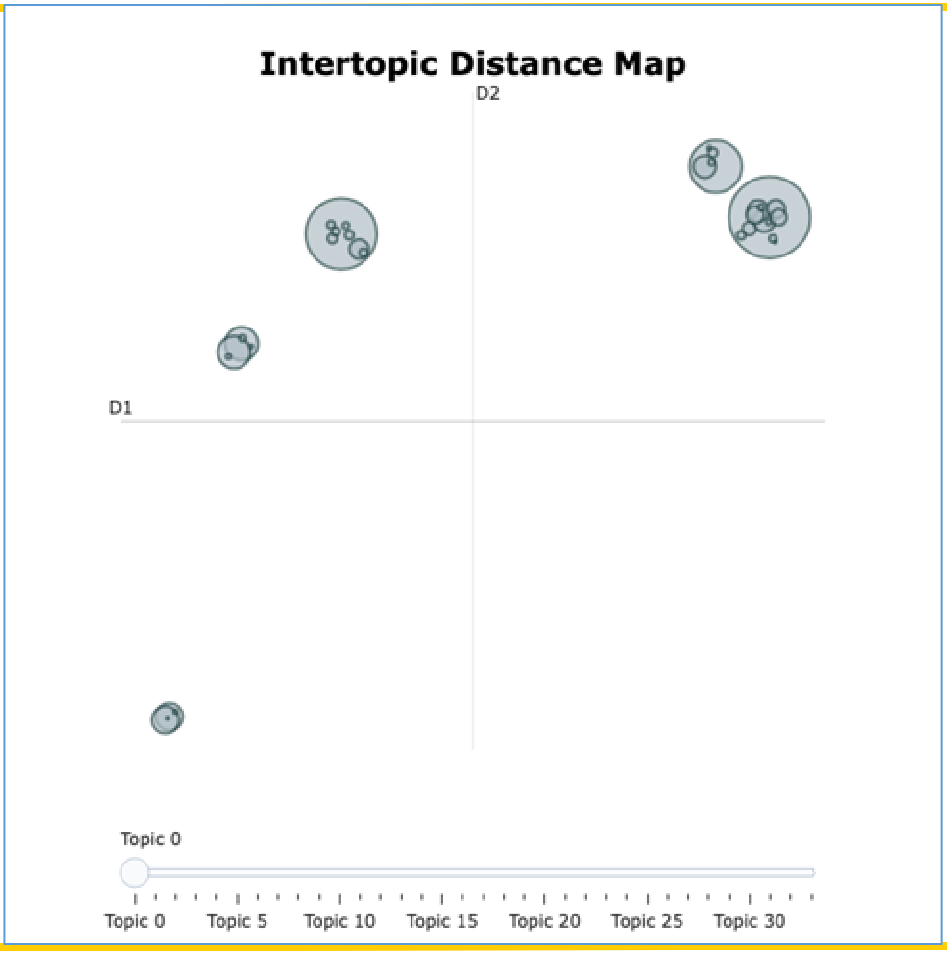
5: Causal Inference and Treatment Effects

6: Bayesian Inference and Probabilistic Models

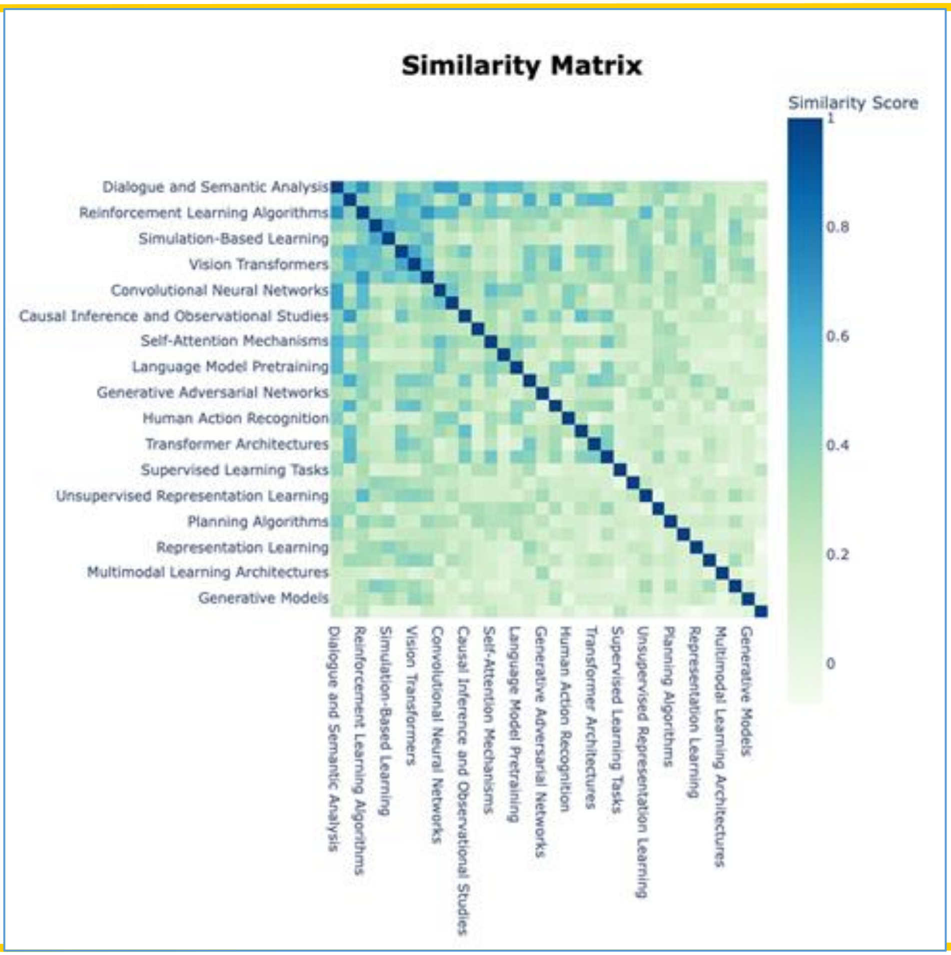
7: Graph Neural Networks

Privacy Notice: Do not record or reproduce.

Relationships between Topics



Relationships between Topics



The Big Picture

- Search for explainability with AI
  - Using LLMs themselves to understand their development
- Possibility of reawakening research in “dead ends”
  - Possible links to the hottest subfields of today
- High level overview of today’s technology to create tomorrow’s

Acknowledgements

- A HUGE thank you to Professor Vivian Lew for her encouragement and advice throughout this entire process, I would not have this project here today without her

- Would also like to thank all of the professors in the Statistics Department for exposing us to the models and algorithms used in research today