# Tom Magelinski

Washington, DC · tmagelinski.github.io · tmagelin@cs.cmu.edu

## Summary of Interests

I build systems that help domain experts understand vast amounts of data through state-of-the-art techniques from natural language processing, generative modeling, graph ML, and network science.

## Education

**Carnegie Mellon University** — Pittsburgh, PA
PhD Computer Science (Societal Computing) *GPA: 3.9* — August 2017 - May 2023
Thesis: "Contextualized Conversational Network Dynamics on Social Media"

**Virginia Tech** — Blacksburg, VA
Honors Baccalaureate Engineering Science and Mechanics *GPA: 3.9* — August 2013 - May 2017
Minors in Math and Physics

**University of Oxford** — Oxford, UK
Visiting Student — January 2015 - April 2015

## Experience

**Johns Hopkins Applied Physics Lab** — Washington, DC
*Senior Data Scientist - Generative AI and Information Extraction* — July 2023 - Present

- Implemented a robust validation framework for an advanced retrieval augmented generation (RAG) pipeline on a large-scale document stream through prompting of large language models
- Partnered in the development of a multi-lingual document clustering service, capable of finding relevant documents to a user's query, clustering documents into stories, and summarizing them
- Researching methods of incorporating multiple media representation techniques such as Whisper and ImageBind to improve upon various multi-modal tasks like any-to-any media search

**Spotify Research** — New York, NY
*Research Scientist Intern* — June 2021 - September 2021

- Improved podcast understanding by integrating heterogeneous social network embedding into a transformer-based pipeline using StellarGraph

**CASOS Lab** — Pittsburgh, PA
*Graduate Research Assistant* — August 2017 - May 2023

- Improved unsupervised multi-modal tweet representational learning through *Deep Tweet Infomax*, which leverages language-aligned word vectors, the conversational graph, hashtags, and URLs. Implemented in PyG and trained on GPU
- Developed highly scalable graph algorithms for bot detection and coordinated actor detection on Twitter datasets with tens of millions of Tweets using igraph and PyTorch
- Built a distributed data analysis pipeline to clean and learn representations of $\sim 100$ million Tweets in PySpark and BigGraph
- Improved SotA graph classification accuracy by 1-2% on social media datasets by creating and implementing a novel deep graph-convolutional architecture in PyTorch
- Built an interactive dashboard in Plotly to analyze Twitter hashtag network dynamics using diachronic node embeddings
- Published in venues like *AAAI*, *TheWebConf*, *ICWSM*, *The Journal of Online Trust and Safety*, *Applied Network Science*, and *IEEE Transactions on Network Science and Engineering*

## Awards and Activities

**Co-Organizer: Ethics for Technologists Lecture Series** — November 2018 - November 2021
**Knight Foundation Fellow** — Fall 2020, Spring 2021
**ARCS Foundation Scholar** — August 2017 - August 2020
**Outstanding Senior: Engineering Science and Mechanics** — May 2017

## Technical Skills

| | |
|---|---|
| Languages and Tools: | Python, Java, R, SQL (MySQL, BigQuery), LaTeX, Git, FastAPI |
| ML Frameworks and Libraries: | PyTorch, TensorFlow, PyG, StellarGraph, HuggingFace, PySpark MLlib |
| Network Science Libraries: | igraph, NetworkX, Graph-Tool |
| Data and Statistics Libraries: | PySpark, NumPy, Pandas, spaCy, NLTK, SciPy, statsmodels, scikit-learn |
| Visualization: | Matplotlib, Seaborn, Plotly |
| Research Interests: | Generative AI, Multi-Modal Learning, Graph Representation Learning, Knowledge Graph Creation, Node Classification, Community Detection, Natural Language Processing |