Alexander S. Tregub

Cell: +1 (330) 671-5352 | GitHub: github.com/AlexTregub | Email: alexs.tregub@gmail.com
Google Scholar: scholar.google.com/citations?user=yvmxP5MAAAAJ

Objectives:

I am a data scientist interested in statistical analysis, computational geometry, developing efficient algorithms, and simulation techniques, with experience in mathematical modeling and computational analysis. As a double major in Applied Mathematics and Computer Science, I have developed my expertise through contributions to multiple research projects, extracurricular projects, and through coursework. As part of Dr. Ruoming Jin's research group, I have applied machine learning and mathematical modeling techniques to projects on classification and language models; my current work is focused on identifying protein-protein interactions for cryogenic electron microscopy data by simulating random geometric configurations of proteins. My other contributions include the development of utilities to process genomic variant data, and of API-querying tools and downstream data processing for NSF-funded projects on data science and digital epidemiology (see preprints listed below). I have contributed to manuscript writing and presented my work in both poster and seminar formats.

Education:

Kent State University: Bachelor's of Science in Applied Mathematics, and Computer Science.
(Expected graduation Spring 2026) **GPA: 3.901**

Honors and Awards:

NASA/OSGC (Ohio Space Grant Consortium) 2025-2026 Fellowship
KSU's SURE (Summer Undergraduate Research Experience) Fellowship (Summer 2023, Summer 2024, Summer 2025)
President's List 2023-2024, Dean's List 2021-2025

Research & Work Experience:

2025 - Present: NASA/OSGC (Ohio Space Grant Consortium) 2025-2026 Fellowship, "Developing Dimensionality Reduction Metric for RNA Editing Data from Microgravity Experiments". Aligned with the mission of the Exploration Systems Development Mission Directorate, Humans in Space.
Mentor: Dr. Jun Li (Kent State University). Project goals: to develop a metric to evaluate differences across RNA editing profiles from large-scale transcriptomics data of cells that experienced microgravity and those that did not, and to evaluate the metric's stability against data perturbations due to data sampling.

2022 - Present: Member of Dr. Ruoming Jin's AI, Machine Learning, and Computational Science Research Lab.

2022 - 2025: Volunteer in Dr. Maimuna Majumder's lab (Boston Children's Hospital), in collaboration with Dr. Helen Piontkivska (Kent State University), contributing as a programmer and data analyst to projects focused on data science, cybersecurity, digital epidemiology, and bioinformatics.

2025: Served as a student leader for the SURE program, assisting with large student meetings, supporting a smaller group of student researchers, and participating in Kent State's summer preview days and sharing my previous experience with the undergraduate research program with visiting prospective students.

2024 - 2025: KSU's SURE (Summer Undergraduate Research Experience) Fellowship (Summer 2024,

Summer 2025), Modeling and simulating protein-protein interactions with an extension of random geometric graphs using random geometric configurations of variable-sized proteins in three dimensions.
Mentor: Dr. Ruoming Jin (Kent State University), in collaboration with Dr. Jack Su (Case Western Reserve University). Presented as a poster at KSU's undergraduate research symposium in Spring 2025, "Modeling Protein-Protein Interactions with an Extension of Random Geometric Graphs".

2023 - 2024: KSU's SURE (Summer Undergraduate Research Experience) Fellowship (Summer 2023), Mentor: Dr. Ruoming Jin (Kent State University). Project goals: to compile existing literature related to optimization on matroids, implement existing algorithms using Python, and explore leveraging the properties of oriented matroids to improve algorithm performance. Presented as a poster at KSU's undergraduate research symposium in Spring 2024, "Analysis of Combinatorial Optimization on Matroids, and their Applications".

Technical Skills:

Programming Languages and Tools: C++ (high-performance modeling, data processing), Python (statistical regression, data visualization, API harvesting), R (visualization), MatLab (building and running models for simple systems), Mathematica (solving complex equations), LaTeX (producing project reports and lab and seminar presentations), Bash (setting up workflows and automation), HTML/CSS (working with web page design), C (experience adding functionality to existing software)
Machine Learning Frameworks: PyTorch, TensorFlow
Virtualization Software: Oracle VirtualBox, VMware Workstation, KVM (setting up virtual machines for courses)
Operating Systems: Unix-based distributions (currently running Arch Linux), Microsoft Windows
Highlighted Coursework: Theory of Statistics, Partial Differential Equations, Design and Analysis of Algorithms, Artificial Intelligence, Big Data Analytics
Strengths: I am a self-driven and self-motivated problem solver who is always ready to embrace challenges and to troubleshoot. I am a team player with a successful track record of working independently and as part of multidisciplinary teams. My ongoing experiences in data analysis for inter-disciplinary projects stimulate my curiosity, adaptability, creativity, communication, and critical thinking skills, and allow me to apply my computational and analytical expertise to real-world problems.

Publications and Preprints:

1. Lubwama, B., Ontiveros, J., Correll Carlyle, R., Kumar, S., Berkane, T., Puri, A., Tregub, A., Nitirahardjo, C., Morgan, E., Lawler, B.C., Aimone, E., H. Piontkivska, and M.S. Majumder. 2025. Practical Considerations for Fine-Tuning BERT-Based Language Models in Health Research: Lessons from Classifying Anti-Vaccine Posts on Social Media. Available at SSRN https://ssrn.com/abstract=5276034. http://dx.doi.org/10.2139/ssrn.5276034
Keywords: natural language processing, text sentiment categorization, fine-tuning, optimizing BERT model performance, benchmarking LLM models
Objectives: Analyzing improvements in BERT language models used for sentiment classification by adjusting training parameters, stopping conditions, model hyperparameters, and the effects of adjusting the fine-tuning dataset properties, by selecting different keywords to be trained from, attempting different labeler conflict resolution approaches, and by collapsing or removing some categories for classification.
Personal Contributions: Developed Python tool utilizing Twitter's APIs to extract tweet text based on a query consisting of relevant metadata and keywords, and then automating the data collection using parallelized Bash scripts. Additionally, created a script for parsing and storing user message data encoded such that both labelers and language models could interpret emojis for additional context for sentiment analysis.

2. Nitirahardjo, C., Morgan, E., Lawler, B.C., Aimone, E., Tregub, A., Puri, A., Ontiveros, J., Correll Carlyle, R., Majumder, M.S., and Piontkivska, H., 2024. Comparing the Usage of Russian-and Ukrainian-Derived Search Terms to Evaluate the Impact of Misinformation, Disinformation, and Propaganda in the US. *Disinformation, and Propaganda in the US (June 20, 2024)*. Available at SSRN: https://ssrn.com/abstract=4871612. http://dx.doi.org/10.2139/ssrn.4871612

Keywords: orthography, search trend analysis, user sentiment analysis, linear regression analysis, bootstrapping

Objectives: Investigating GST (Google Search Trend) data separated by US state and utilizing available user demographic metrics to gauge users' attitudes towards certain terms using their spelling in search queries. Accomplishing this by performing in-depth univariate and multivariate linear regression per state per metric, and highlighting relevant findings. Additionally, performing random resampling analysis (bootstrapping) to ensure trends exist due to genuine user sentiment and not flaws in search trend data.

Personal Contributions: Developed Python tool to perform automated resampling and regression analysis on search trend data, and created visualizations from the produced data for interim results.

3. Ontiveros, J., Correll Carlyle, R., Puri, A., Kumar, S., Tregub, A., Nitirahardjo, C., Morgan, E., Lawler, B.C., Aimone, E., Piontkivska, H., and Majumder, M.S., 2023. Classification Performance Thresholds for BERT-Based Models on COVID-19 Twitter Misinformation. Available at SSRN 4489865. http://dx.doi.org/10.2139/ssrn.4489865

Keywords: training data selection, logistic linear classifier, NLP, distilling, fine tuning, web scraping, data labeling, social media sentiment analysis

Objectives: Examining BERT models for sentiment classification, based on previous research detailing sentiment analysis on Twitter data, using a newly collected dataset of tweets collected based on predetermined keywords. Then, comparing our fine-tuned BERT models against simpler Logistic models and reviewing their potential issues.

Personal Contributions: Developed Python tool leveraging Twitter's API to extract tweet IDs based on text queries, automating the collection of IDs to be used with the Hydrator tool to extract their full text.

Other experiences:

- Developed C++ utility for processing and merging large binary alignment map (BAM) datasets, 2025.
- Developed utilities for parsing variant calling (VCF) and binary alignment map (BAM) data in Python and C++, 2023 – 2025.
- Assisted with website migration for Dr. Leanne Powner in HTML and CSS, 2022. https://www.leannecpowner.com/
- 2nd Dan Black Belt with Kwanmukan, with Mr. Patrick M. Hickey's Dojo, gaining leadership experience from assisting with classes.