Yueran (Hannah) Sun

Master's Student in Data Science | University of Washington | Seattle, WA

yuerans@uw.edu | LinkedIn | GitHub | Personal Website

Education

University of Washington **Seattle, WA**
Master of Science in Data Science, GPA: 4.0/4.0 **Sept 2025 - Present (Expected Mar 2027)**
Relevant Coursework: Applied Statistics and Experimental Design, Software Design, Data Visualization, Statistics and Probability

University of Michigan **Ann Arbor, MI**
Bachelor of Science **Aug 2021 - May 2025**
Major: Data Science, Minor: Mathematics, GPA: 3.71/4.0 (University Honors)
Relevant Coursework: Natural Language Processing, Machine Learning, Data Mining, Database Management, Web Systems, User Interface Development, Computation Theory, Discrete Math, Financial Math, Differential Equations, Multivariable Calculus, Matrix Algebra

Research Experience

University of Washington | Lab for Computing Cultural Heritage [Link] **Nov 2025 - Present**
Graduate Student Researcher | Mentor: Prof. Benjamin Charles Germain Lee Seattle, WA

- Examined clusters from over 1.5 million Library of Congress archival newspaper images to identify large-scale patterns of visual reuse and circulation in early 20th-century print media.
- Applied CLIP embeddings with DBSCAN clustering to generate groupings of visually similar images, then computed cluster statistics to support systematic interpretation.
- Built embedding-image lookup pipelines linking CLIP vectors, image IDs, filenames, and metadata, enabling reliable retrieval and scalable downstream analysis.
- Performed zero-shot concept labeling on clusters by embedding candidate semantic categories and assigning top concepts using cosine similarity.
- Quantified image circulation and reuse dynamics by measuring cluster frequency distributions, cross-newspaper diversity, and temporal spread.

University of Washington | Language Accessibility Research Lab [Link] **Sept 2025 - Present**
Research Collaborator | Primary Collaborator: Anukriti Kumar | PI: Prof. Lucy Lu Wang Seattle, WA

- Contributed to NeuroAdapt, a content personalization framework across neurodivergent profiles, by developing core segmentation, alignment, and post-processing pipelines comparing original text with plain-language counterparts.
- Built a semantic alignment pipeline using LASER-3 embeddings and VecAlign to support varying alignment types and capture sentence splitting, merging, expansions, and omissions.
- Conducted quantitative evaluation using linguistic and readability metrics, complemented by qualitative analysis, to assess lexical substitutions, structural changes, semantic restorations, and tone and style adjustments.
- Designed a controlled evaluation study protocol to rigorously assess whether productive friction enhances deep comprehension and long-term retention in adaptive reading systems.

MobiDrop (Zhejiang) Co., Ltd **Nov 2023 - May 2024**
Bioinformatics Research Assistant | Mentor: Dr. Guantao Zheng Shanghai, China (Remote)

- Pretrained scGPT, a Transformer-based model for gene expression prediction using single-cell RNA sequencing data.
- Curated and preprocessed a dataset of 300,000 human blood cells from the CELLxGENE repository, preserving organism-level structure while performing normalization, binning, and tokenization of highly variable genes.
- Modified the Transformer encoder architecture to jointly embed gene identities and binned expression values.
- Leveraged GPU-accelerated training to scale pretraining experiments, writing Bash scripts to automate job submission, checkpointing, and experiment tracking.

Industry Experience

Develop for Good **Oct 2025 - Feb 2026**
Product Manager, PainUSA Seattle, WA (Remote)

- Led a team of five undergraduate students to design and deliver the PainUSA non-profit website, coordinating milestones, delegating technical tasks, and aligning design and client requirements.
- Guided user interface prototyping in Figma and implementation of the public website in Webflow.
- Engineered and deployed a Mapbox-based interactive clinician lookup map, leveraging Cloudflare R2 for data storage and hosting, and embedded HubSpot forms for newsletter sign-up.
- Managed client communication with Stanford Division of Pain Medicine leadership, producing structured handoff documentation to ensure sustainable post-delivery ownership.

Ternary **Jun 2025 - Aug 2025**
Software Engineer Intern, Product Delivery Team | Supervisor: Kyle Rattet Seattle, WA (Remote)

- Upgraded Ternary's cloud cost forecasting API by integrating Meta Prophet, enabling automated daily and weekly client spend predictions and improving modeling sophistication beyond the prior linear regression baseline.
- Improved forecast reliability through time-series cross-validation, hyperparameter tuning, and confidence interval integration.
- Engineered backend workflows connecting Go and Python services, implementing schema-validation tests and CI pipelines to ensure forecasting accuracy and production stability.
- Deployed and maintained staging environments on Google Cloud, using Terraform to provision and manage shared resources and support pre-production UI testing.
- Enhanced the frontend user interface with a rotating slideshow of interactive line charts, visualizing top cost-contributing business dimensions.

University of Michigan | Multidisciplinary Design Program [Link] **Jan 2024 - Dec 2024**
Student Developer, ProQuest Team | Mentor: Prof. Sindhu Kutty Ann Arbor, MI

- Automated textual and layout segmentation of historical Detroit Free Press front pages, extracting article-level structure such as titles, bylines, body text, and reading order from JP2 images and OCR data.
- Developed an MLOps pipeline integrating large language models, Gaussian mixture models, and computer vision for text-type classification, article boundary detection, and noise labeling.

- Validated segmentation and classification performance through controlled A/B testing, comparing pipeline variants to optimize accuracy, robustness, and processing efficiency.
- Reduced segmentation cost by 50% compared to manual labeling while increasing throughput to 12 pages per minute.

Pachira Information Technology (Hengqin) Co., Ltd **May 2024 - Aug 2024**
NLP Algorithm Intern | Supervisor: Dr. Chenglong Ma Zhuhai, China

- Enhanced the retrieval augmented generation pipeline powering an in-car voice assistant for Toyota vehicles.
- Developed a new warning-light query handling module and deployed multiple system iterations through vehicle simulator testing.
- Implemented multi-intent recognition, refusal logic, and query translation components to improve accuracy, safety, and user interaction quality.

Publications

- Kumar, A., Glazko, K. S., Sun, Y., Harniss, M., Wang, L. L., & Mankoff, J. *Beyond Readability Metrics: Plain Language Priorities in Disability Advocacy Organizations.* Accepted to ACM FAccT 2026.
- Lee, B. C. G., Buccalon, B., & Sun, Y. *Viral Images: Identifying Re-printings within 1.5 Million Photographs in Chronicling America.* Submitted to ACM Journal on Computing and Cultural Heritage, Special Issue on Visual Heritage.

Skills

- **Languages:** Python, R, SQL, C/C++, Go, JavaScript/TypeScript, Bash, HTML/CSS
- **Tools:** Git, Linux, Google Cloud Platform (GCP), AWS, Docker, Terraform, MongoDB, IBM Db2, DBeaver, Tableau, Jupyter, Conda, GitHub Actions (CI/CD)
- **Frameworks & Libraries:** PyTorch, scikit-learn, XGBoost, LightGBM, LangChain, FastAPI, React, Node.js, Pandas, NumPy, Dask, OpenCV