

Marko J. Sterbentz

LinkedIn: [linkedin.com/in/MarkoSterbentz](https://www.linkedin.com/in/MarkoSterbentz) ♦ GitHub: github.com/MarkoSterbentz ♦ Personal: [MarkoSterbentz.com](https://www.MarkoSterbentz.com)

EDUCATION

Northwestern University Ph.D. Computer Science, GPA 4.0/4.0	Evanston, IL Sep. 2019 – Present
University of Southern California M.S. Computer Science, GPA 3.83/4.0	Los Angeles, CA May 2019
Idaho State University B.S. Computer Science with Mathematics minor, GPA 3.99/4.0	Pocatello, ID May 2017

WORK and RESEARCH EXPERIENCE

Northwestern University <i>Research Assistant – advised by Dr. Kristian Hammond</i>	633 Clark St, Evanston, IL 60208 Sep. 2019 – Present
<ul style="list-style-type: none">• Developing methods for performing analytics augmented generation with LLMs to produce contextualized answers derived from large-scale relational databases.• Researching automated approaches for generating high-quality, in-domain text-to-SQL training data.• Exploring novel uses of LLMs for planning and reasoning about data science processes, enabling the generation of data analysis reports that emulate human analyst workflows.	
Lawrence Livermore National Laboratory (LLNL) <i>Research Intern</i>	700 East Ave, Livermore, CA 94550 May 2019 – Aug. 2019
<ul style="list-style-type: none">• Developed a new software component using C++ and Python for performing material interface reconstruction.• Integrated code into LLNL's open-source HPC framework Axom using best software engineering practices.• This project is open source and the code is available on GitHub.	
Idaho National Laboratory (INL) <i>Research Intern</i>	2525 Fremont Ave, Idaho Falls, ID 83415 May. 2018 – Aug. 2018
<ul style="list-style-type: none">• Enhanced an INL volume visualization system using C#, HLSL, and compute shaders in the Unity game engine.• Utilized raymarching in tandem with a specialized data format to enable interactive visualization of exascale data in immersive environments.• Presented associated research paper at PEARC18 conference in July 2018.	
<i>Research Intern</i>	May. 2017 – Aug. 2017
<ul style="list-style-type: none">• Developed software using the Unity game engine in C#, HLSL, and compute shaders for real-time rendering of exascale volume data for use in immersive virtual reality environments and conventional desktops.• Collaborated with lab researchers to ensure this software would satisfy their use cases and practical requirements.	
<i>Software Engineering Intern</i>	May. 2016 – Aug. 2016
<ul style="list-style-type: none">• Built software components in Java and C++ for a large-scale data streaming and rendering platform.• Coordinated with other developers using agile development techniques and git.• Conducted in-depth presentations/demonstrations of the INL's computer aided virtual environment (CAVE) 3-D visualization capabilities onsite and in local classrooms.	

Software Engineering Intern

May. 2015 – Aug. 2015

- Generated improved interfaces using C++ and the Virtual Reality User Interface API for immersive visualization software.
- Aided in setup of remote collaboration tools to be utilized by researchers using the CAVE.

Software Engineering Intern

May. 2014 – Aug. 2014

- Created a new immersive visualization application in C++ that was capable of rendering both LiDAR and 3-D models simultaneously.
- Implemented additional control features, basic animations, scaling, and positioning of models.

Idaho State University

921 S 8th Ave, Pocatello, ID 83209

Undergraduate Research Intern

Nov. 2015 – May 2017

- Wrote software in C++ that interfaced with the Velodyne VLP-16 LiDAR sensor, extracted the useful information from incoming data packets, and registered data points using an iterative closest point algorithm.
- Added functionality to extract data from an inertial measurement unit (IMU), send it over a wireless network, and recreate the scanned environment on the user's laptop in real time.
- Constructed initial plan to meet the project goals in terms of hardware, software, and output required.
- Work performed as part of a study to determine the state of plant life in Idaho utilizing unmanned aerial vehicle LiDAR data.
- Funded by NSF / Idaho EPSCoR as part of the MILES Undergraduate Research Internship Program.

RECENT PROJECTS

Satyrn: A Platform for Analytics Augmented Generation *(under development)*

- Creating a system that leverages knowledge of analytic processes to derive information from relational databases and communicate the results in fluent and coherent natural language reports.
- Implements a method similar to retrieval augmented generation (RAG) for retrieving and analyzing data, which is then processed by a large language model (LLM) to produce factual, fluent, and coherent language output.
- Generates more factually accurate reports than GPT-4 Code Interpreter using smaller models like Mistral-7B.
- GitHub Repository: <https://github.com/nu-c3lab/satyrn>
- Paper: <https://arxiv.org/abs/2406.12069>

Automating the Generation of In-Domain Text-to-SQL Training Data *(under development)*

- Developing a system that enables users to define schema-independent plan templates for generating diverse question-SQL pairs across any database, supporting text-to-SQL training paradigms.
- Streamlines data generation by enabling users to create a single template for a specific question type, which can be applied across all datasets without incorporating schema details like joins.
- Improves efficiency by eliminating the need for users to configure new templates for each database, ensuring high-quality text-to-SQL data generation with minimal effort.

Mim-IR: Open-Domain Question Answering over Large Scale Text Corpora

- Created a system for answering complex questions via information retrieval and analysis over Wikipedia articles.
- Trained large scale language models for question answering and parsing questions into structured representations.
- Implemented a suite of analytic components that utilize the structured representation and intermediate results for answering questions which require numeric, set, and logical reasoning.
- GitHub Repository: <https://github.com/nu-c3lab/mim-ir>

Classifying Police Complaint Reports

- Undertook a deep dive into data consisting of complaints filed against the Chicago Police Department from the Invisible Institute's Citizens Police Data Project.
- Trained a language model which achieved an 84% accuracy rate to classify the text of each police report as belonging to one of multiple categories of complaints.

- Using the trained model, it was found that a disproportionate number of unclassified reports belonged to the “use of force” category of complaints.

COVID Resource Response

- Collaborated with fellow lab members and corporate partners at Rheaply and Kin + Carta to build a website and application to facilitate the exchange of personal protective equipment and other resources needed to prevent the spread of COVID-19 in the Chicago area.
- Designed and implemented a database schema to support the storage and access needs of this application and developed Python code using SQLAlchemy to connect the database to the frontend web interface.
- Assisted in the development of a dashboard for visualizing COVID cases with an interactive nationwide map that automatically pulled in the most up-to-date statistics on a daily basis.

PUBLICATIONS and PREPRINTS

Marko Sterbentz, Cameron Barrie, Shubham Shahi, Abhratanu Dutta, Donna Hooshmand, Harper Pack, and Kristian Hammond. “Satyrn: A Platform for Analytics Augmented Generation.” To appear at the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024 Main). Preprint available at: <https://arxiv.org/abs/2406.12069>

David Demeter, Oshin Agarwal, Simon Ben Igeri, **Marko Sterbentz**, Neil Molino, John Conroy, and Ani Nenkova. “Multi-domain Summarization from Leaderboards to Practice: Re-examining Automatic and Human Evaluation.” In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* at EMNLP 2023.

Marko Sterbentz, Cameron Barrie, Donna Hooshmand, Shubham Shahi, Abhratanu Dutta, Harper Pack, Andong Li Zhao, Andrew Paley, Alexander Einarsson, and Kristian Hammond. “Lightweight Knowledge Representations for Automating Data Analysis.” On *Arxiv*: <https://arxiv.org/abs/2311.12848>

Andong L. Li Zhao, Andrew R. Paley, Rachel F. Adler, Harper Pack, Sergio Servantez, Alexander Einarsson, Cameron Barrie, **Marko Sterbentz**, and Kristian Hammond. “Requirements for Open Political Information: Transparency Beyond Open Data.” In *AAAI FSS-21 Artificial Intelligence in Government and Public Sector*. 2021.

Andrew R. Paley, Andong L. Li Zhao, Harper Pack, Sergio Servantez, Rachel F. Adler, **Marko Sterbentz**, Adam Pah, David Schwartz, Cameron Barrie, Alexander Einarsson, and Kristian Hammond. “From Data to Information: Automating Data Science to Explore the U.S. Court System.” In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021.

James H. Money, **Marko Sterbentz**, Nathan Morrical, Thomas Szewczyk, and Landon Woolley. “GPGPU Enabled Ray Directed Adaptive Volume Visualization for High Density Scans.” In *Proceedings of the Practice and Experience on Advanced Research Computing*. 2018.

CONFERENCES, POSTER SESSIONS, and PRESENTATIONS

M. Sterbentz, K. Weiss. Improving Multi-Material Simulations: A Material Interface Reconstruction Component in Axom. *Lawrence Livermore National Laboratory Intern Expo and Poster Session*. Livermore, CA, August 2019.

M. Sterbentz, J. Money. GPGPU Enabled Adaptive Volume Visualization Using Commodity Game Engines. *Idaho National Laboratory Intern Expo and Poster Session*. Idaho Falls, ID, August 2018.

M. Sterbentz. Large Scale Adaptive Volume Visualization Using GPGPU Techniques and Commodity Game Engines. *Idaho National Laboratory Technical Presentation*. Idaho Falls, ID, July 2018.

M. Sterbentz, M. Johnson, A. Syal, R. Chugh, P. Taneja, J. Tang. Prismo: An Affective Computing Platform Built for Microsoft Hololens. *USC Games Expo 2018*. Los Angeles, CA, May 2018.

M. Sterbentz, J. Money. Adaptive Volume Rendering for Exascale Data Using Immersive Environments. *Idaho National Laboratory Intern Expo and Poster Session*. Idaho Falls, ID, August 2017.

G. Cochrane, M. Sterbentz, J. Edwards. Real-Time LiDAR Terrain Mapping and Analysis. *ISU Undergraduate Research Symposium 2017*. Pocatello, ID, April 2017.

G. Cochrane, M. Sterbentz, J. Edwards. Real-Time LiDAR Terrain Mapping and Analysis. *Idaho EPSCoR Annual Meeting 2016*. Coeur d'Alene, ID, October 2016.

M. Sterbentz. Enhancing Scientific Research with Virtual Reality. *Math/CS Club Science, Math, Engineering, and Related Fields (SMERF) Talks*. Pocatello, ID, October 2016.

M. Sterbentz, J. Money. LIVE2: An Engine for Dynamic and Distributed Visualization. *Idaho National Laboratory Intern Expo and Poster Session*. Idaho Falls, ID, August 2016.

G. Cochrane, M. Sterbentz, J. Edwards. Real-Time LiDAR Terrain Mapping and Analysis. *Idaho Conference on Undergraduate Research (ICUR)*. Boise, ID, July 2016.

M. Sterbentz, E. Whiting. Dynamic Data Manipulation in the CAVE. *Idaho National Laboratory Intern Expo and Poster Session*. Idaho Falls, ID, August 2015.

M. Sterbentz, E. Whiting. Building a Hybrid Model Viewer to Enhance the Capabilities of the Computer Assisted Virtual Environment. *Idaho National Laboratory Intern Expo and Poster Session*. Idaho Falls, ID, August 2014.

AWARDS, HONORS, and GRANTS

2021 – 18th Intl Conf on Artificial Intelligence and Law: Peter Jackson Award for Best Innovative Application Paper

2017 – University of Southern California Viterbi Dean's Scholarship

2017 – Idaho State University College of Science and Engineering High Honors Designation

2016 – MURI Program and research funding award sponsored by the National Science Foundation/Idaho EPSCoR

2016 – Google IgniteCS Grant for community mentorship program

2015 – MURI Program and research funding award sponsored by the National Science Foundation/Idaho EPSCoR

2014 – Idaho National Laboratory Intern Poster Session, Overall, 2nd Place Award

2014 – Idaho National Laboratory Intern Poster Session, Best Oral Presentation, 2nd Place Award

2014 – Center for Advanced Energy Studies (CAES) Energy Scholar Award

Idaho State University College of Science and Engineering Dean's List – 8 of 8 semesters

2013 – Idaho National Laboratory Scholarship Recipient – top tier

2013 – Idaho State University Presidential Scholarship

VOLUNTEER EXPERIENCE

Responsible AI Student Organization (RAISO)

Evanston, IL

Mentor

Jan. 2022 – June 2022

- Mentored Northwestern undergraduate students interested in data science and machine learning.

CS PhD Advisory Council Buddy Program

Evanston, IL

Peer Mentor

Sept. 2021 – Aug. 2022

- Mentored and advised two new PhD students in Northwestern University's CS department.

Viterbi Graduate Mentorship Program

Los Angeles, CA

Peer Mentor

June 2018 – May 2019

- Mentored and advised two new graduate students in the USC Viterbi School of Engineering's CS department.

Google IgniteCS

Pocatello, ID

Program Mentor

Aug. 2016 – April 2017

- Co-wrote the initial grant proposal to Google and obtained funding for a mentorship program to teach children from underrepresented groups in computer science how to write code.

- Instructed high school students in basic programming techniques through the use of the Scratch programming language and hosted an additional coding workshop for local elementary school students.

TEACHING EXPERIENCE

Northwestern University

Teaching Assistant – CS 338 (Practicum in Intelligent Information Systems)

Evanston, IL

Sept. 2024 – Dec. 2024

Sept. 2022 – Dec. 2022

Jan. 2022 – June 2022

Sept. 2020 – Dec. 2020

University of Southern California

Teaching Assistant / Course Producer – CSCI 576 (Multimedia Systems Design)

Los Angeles, CA

Aug. 2018 – May 2019

SKILLS and PROFICIENCIES

Programming Languages: Python, SQL, C/C++, Java, C#, Javascript, Lisp

Frameworks and Tools: Huggingface Transformers, Scikit-Learn, SpaCy, NLTK, Elasticsearch, Pytorch, SQLAlchemy

RELEVANT COURSEWORK

Graduate

Machine Learning, Data Science Seminar, Foundations of Artificial Intelligence, Deep Learning Foundations from Scratch, Natural Language Processing, Seminar in Statistical Language Modeling, Database Systems, Multimedia Systems Design, Computer Animation and Simulation, Digital Geometry Processing, Knowledge Representation and Reasoning, Conversational Interfaces, Computational Creativity, Affective Computing

Undergraduate

Statistical Methods, Introduction to Probability, Differential Equations, Linear Algebra, Database Design and Implementation, Data Structures and Algorithms, Advanced Computer Graphics

LEADERSHIP and PROFESSIONAL AFFILIATIONS

2018 – Present: Member of Association for the Advancement of Artificial Intelligence (AAAI)

2013 – Present: Member of Association for Computing Machinery (ACM)

2016 – 2017: President of the ISU Math/CS Club

2014 – 2016: Secretary of the ISU Math/CS Club

2013 – 2014: Secretary of the ISU Green-Up Club

ACADEMIC SERVICE

2022-2024: Admission committee member for Northwestern University's Master of Artificial Intelligence program

2023-2024: Overseeing and advising MSCS, MSAI, and undergraduate CS students on AI research projects.

2020: Reviewer for Neural Information Processing Systems (NeurIPS) Conference

2020: Supervised 27 Northwestern computer science undergraduates on machine learning and data science projects

2018: Session Chair of Data Analytics / Deep Learning Session at Practice and Experience on Advanced Research Computing '18 Conference. Pittsburgh, PA. July 22 – 26, 2018.