Zach Hafen-Saavedra

z.hafen.saavedra@gmail.com || zhafen.github.io || (303) 819-8840 || Chicago, IL ||

Summary

[click here for a work sample]

Always-growing scientist with over 10 years of experience leading efficient solution development for complex problems, including 10 years of Python experience, 8 years profiling the histories of individual entities (i.e. customer analogs), and 4 years using natural-language processing (NLP) to predict academic profit.

Education

The Erdős Institute 2023
Data Science Certificate Irvine, CA

Northwestern University 2020
PhD, MS, Physics and Astronomy Evanston, IL
Specialization: Astrophysical Data Analysis

University of Northern Colorado 2014
BS, Mathematical Physics Greeley, CO

Experience

Far Horizons Data Scientist September 2023 - Present
Adler Planetarium Chicago, IL

- Utilized industry-standard tools including Docker, AWS CodeBuild, Amazon ECR, Amazon S3, and PostgreSQL to deploy an end-to-end data analytics pipeline on AWS, enabling non-technical stakeholders to ingest, process, and export 100s of GB of data via a user-friendly web interface.
- Developed a Python-centric computer-vision pipeline to automate the process of aligning nighttime aerial images (only localized to within 5 km) with daytime images, dramatically increasing the alignment rate from 4 images/hour to 5000 images/hour.
- Developed documentation, an intuitive user interface, a suite of 40+ code tests, and a stable, containerized computing environment, preparing for 3 years of minimal-maintenance use by stakeholders.
- Directed the adoption of Agile project management for museum volunteers, seamlessly integrating the workflow for DevOps volunteers, mechanical-engineer volunteers, and Adler staff.
- As a museum resident scientist, educated and collaborated with non-technical educators to deliver life-changing deep-impact programs for 20+ high-school students and 4 interns.

Other June 2023 - Present

- Trained a convolutional neural network to perform sentiment analysis of cat meows (i.e. time-series audio data), achieving 90% validation accuracy and earning a merit in the Erdős Data Science program.
- Fine-tuned the Llama LLM with RLHF (reinforcement learning from human feedback) to, given an input paper, summarize the 100 most-textually-similar abstracts.
- Mentored a UCI computational-linguistics PhD student, leading to a presentation at AI4Science on our ongoing NLP academic-profit work and the imminent completion of our related paper.

Business Data Analyst June 2023 - September 2023
Northwestern University, Center for Interdisc. Explor. and Research in Astrophysics Evanston, IL

- Created a secure, web-based business-intelligence dashboard using Streamlit, enabling business staff to analyze data and present updated, tailored visualizations to stakeholders.
- Implemented a low-dependency shell-scripting+Python solution stack tailored to organization resources, guaranteeing operational continuity and maintainability of solutions.
- Updated the BI workflow to automatically ingest financial data sourced from Cognos BI, restoring secure access to crucial financial information.

- Automated PDF text mining and identified key data-centric insights, empowering stakeholders to form a six-point evidence-based action plan for improving diversity, equity, and inclusion.

McCue Prize Postdoctoral Fellow in Cosmology July 2020 - June 2023
University of California–Irvine, Department of Physics and Astronomy Irvine, CA

- Developed a Python-frontend, C++-backend NLP data pipeline to perform embeddings of scientific text (e.g. Word2Vec), and identified clustering-related metrics correlated with a 150% increase in academic profit (i.e. citations).
- Utilized ML tools (inc. sklearn, PyTorch) to construct an ensemble voting model for paper impact as measured by number of citations, with a validation root-mean-square error 2/3$\times$ of the baseline error.
- Automated data retrieval from NASA APIs, extracting abstracts, references, citations, etc. for 1M+ papers and generating a collaborator-ready vector database.
- Orchestrated a mock data challenge spanning nine international institutions, quantifying parameters for statistical models to attain 90%+ accuracy in estimating the chemical composition of intergalactic gas, subsequently informing technical stakeholders' modeling decisions.
- Incorporated a new data source into an open-source Git repository, decreasing mean parameter-estimation error by up to half.
- Led and organized a workshop of twenty key galaxy-community leaders, fostering cross-specialty dialogue to discern high-value research targets and leading to the development and completion of 3+ projects.
- Built an end-to-end work flow linking three disparate sources of structured and semi-structured data (black hole, star cluster, and galaxy simulations) and predicting focus areas for stakeholders in adjacent disciplines (gravitational wave astrophysics).

National Science Foundation Graduate Fellow in K-12 Education June 2014 - July 2020
Northwestern University, Department of Physics and Astronomy Evanston, IL

- Used remote high-performance-computing resources to apply probabilistic and exact matching to 20+ TB of relational data, reducing to 50 GB of event-tracking data and isolating parameters stakeholders could use to predict future behavior with 99%+ certainty.
- Performed time-series decision-tree classification to predict the cosmic origins of the atoms we are made of, delivering concrete, clear, and testable hypotheses to guide collaborators.
- Employed software-development best practices such as version control, code review, testing, and continuous integration to develop 13+ open-source packages and contribute to 6 more.
- Responsibly used remote resources to operate 100,000-CPU-hour simulations, generating vector databases used by 30+ stakeholders to increase statistical power and realism.
- Crafted award-winning visualizations displayed throughout Chicago libraries and museums–including the Museum of Science and Industry–advertising the beauty of science to a wide audience.
- Partnered with the Northwestern Academy for Chicago Public Schools to pioneer a high-school data-science education program, reaching over 100 underrepresented students from across Chicago.
- Founded the Physics Graduate Student Council to improve student life, retention, and recruitement, tied to a nearly 200% increase in student recruitement.
- Collaborated with 100+ researchers, leading to 36 published papers, 7 as lead author.

Skills

Techniques: machine learning, NLP, neural networks, statistics (frequentist/Bayesian), data analysis (inc. cleaning, visualization, warehousing), containerization, dashboarding, code testing/CI, GIS, computer vision
Interpersonal skills: technical leadership and management, storytelling, mentoring
Tools: Python (inc. pandas, scikit-learn, pytorch), AWS (inc. S3, end-to-end pipelines), PostgreSQL, Docker, BI (Cognos BI, Streamlit), NoSQL, parallel computing, git, C/C++, GDAL, OpenCV, Windows/Mac/Unix