Zachary Webel

Data Scientist & Applied Mathematician

Maryland Ave, Washington D.C

Washington, D.C.

561-225-4251

zow2@georgetown.edu

@zac-webel

zacharywebel

Profile

Skilled engineer with over 8+ years project experience with Python, 5+ years working with Machine and Deep Learning and experience with TensorFlow, Keras, Pandas, Numpy, Scikit Learn, PyTorch, Matplotlib, ggplot, Tableau, Gensim, NLTK, NEO4J, PostgreSQL, R, C++, Docker, CUDA, MLOps, Generative AI, Large Language Models (LLM), BigData, Google Cloud and AWS. Continuously staying updated with the latest industry trends and advancements, I am **committed to applying cutting-edge methodologies to deliver impactful solutions that address complex business challenges**.

Experienced Data Scientist and Applied Mathematician with a **passion for leveraging analytical skills to extract valuable insights from complex datasets**. Possessing a strong background in mathematics and statistics, coupled with expertise in machine learning and data mining techniques, I am adept at transforming raw data into actionable solutions. A **collaborative and innovative problem-solver**, I thrive in multidisciplinary teams, effectively communicating technical concepts to non-technical stakeholders.

Education

2022 – 2023

Master of Science: Health Informatics & Data Science
Georgetown University – Washington, D.C.
GPA: 4.0

Courses Taken:

- Machine Learning
- EHR Data Science
- Genomic Data Science
- Imaging Artificial Intelligence
- Advanced Health Informatics
- AI for Health Applications
- Digital Health Application Design
- Biostatistics
- Research Capstone

2019 – 2022

Bachelor of Science: Applied & Computational Mathematics
Florida State University Honors College – Tallahassee, FL.
GPA: 3.9 – Magna Cum Laude

Courses Taken:

- Ordinary Differential Equations
- Linear Algebra
- Partial Differential Equations
- Numerical Analysis
- Math Modeling
- Complex Variables
- Discrete Mathematics
- Vector Calculus
- Calculus I, II, and III
- Economics
- Statistics
- Physics A & B

2019

Shanghai Jiao Tong University: Machine Learning
Shanghai, China

.

I was a member of the Machine Learning immersion program at the Shanghai Jiao Tong University in Shanghai, China. While in the program, I took a course in Machine Learning which provided me with a deeper understanding of the theoretical foundations and practical applications of Artificial Intelligence. I also attended China Shanghai International Technology Fair.

Experience

05 2023 – 08 2023

Data Scientist - Artificial Intelligence Researcher – Large Language Models
Food and Drug Administration (FDA), Washington, D.C.
Repository: https://github.com/zac-webel/influenza-11M

As an Artificial Intelligence Researcher at the FDA, I built a nucleotide transformer language model from scratch to predict mutations in virus RNA coding for glycoproteins. My model was able to simulate the evolution of the virus genome to provide insight to guide viral countermeasures towards future versions of the virus before the virus has evolved to that stage.

- Built 11M parameter LLM from scratch using TensorFlow Keras.
- User inputs RNA sequence and model generates a predicted evolved sequence with substitiutions, insertions and deletions.
- Python, Keras, Pandas
- Code: https://github.com/zac-webel/influenza-11M/blob/main/influenza-11m-notebook.ipynb
- Paper: https://github.com/zac-webel/influenza-11M/blob/main/capstone_influenza_11M_final_paper.pdf

2023

Generative Artificial Intelligence Research – Language Model Virus Protein Design
Georgetown University
Repository: https://github.com/zac-webel/Tail-Fiber-LSTM

Created a sequential language model that generates novel virus tail fiber protein sequences.

- Created a two branched neural network which shows the ability to generate a novel protein sequence with a third party (ProteInfer) predicted function of virus tail fiber and cell binding when initialized with a random seed sequence and user desired molecular properties.
- User inputs 30 amino acid seed and desired molecular properties for the sequence, the model, at each time steps generates a 21-class probability distribution representing 20 amino acids and a stop flag.
- Created script to pull and clean protein sequences of interest from the NCBI Entrez protein database.
- Tokenized all sequences and performed word2vec to generate custom amino acid embeddings.
- Wrote a vector embedding translation script.
- At each time step three molecular properties vectors are concatenated: target properties, window properties and difference properties.
- Created, trained, and evaluated own model architecture containing only 4 million parameters.
- Wrote a generation script.
- Model Code: https://github.com/zac-webel/Tail-Fiber-LSTM/blob/main/TAIL_FIBER_STEP_5.py
- Paper: https://github.com/zac-webel/Ai-Final-Paper/blob/main/Protein_Paper.pdf

2023

Generative Artificial Intelligence Research – Language Model Polypeptide Secondary Structure Design
Georgetown University
Repository: https://github.com/zac-webel/Prote-Fold

SELF ATTENTION – LSTM HYBRID GENERATES PROTEIN SEQUENCES WITH DESIRED SECONDARY STRUCTURE PROPERTIES

- Created a neural network which takes input a seed sequence and 3 desired secondary structure percentages and is shown to construct novel polypeptide sequences with an average absolute percentage difference of under 3% for alpha helix, turn and beta sheet desired properties calculated by bio seqUtils protein analysis.
- Created a data collection script from the NCBI Entrez protein database.
- Tokenized all sequences and performed word2vec to generate custom amino acid embeddings.
- Wrote a vector translation script.
- At each time step three molecular properties vectors are concatenated: Target properties, window properties and difference properties all of which contain 3 percentages corresponding to helix, turn and sheet.
- Created, trained, and evaluated model architecture containing less than 6 million parameters.
- Wrote a generation script.
- Model Code: https://github.com/zac-webel/Prote-Fold/blob/main/PROTE_FOLD_STEP_5.py

05 – 07 2023

NLP Engineer – Data Scientist – Private Intelligence
Stealth Startup, Washington, D.C.

- Deployed 3b parameter language model summarizing scientific papers.
- Deployed several pretrained models from Huggingface – summarization, ner, token classification.
- Created a dynamic topic model to model the evolution of research topics in a target country.
- Created a knowledge graph of foreign institutions/scientists collaborating on technology advancements.
- Created a directed word embedding graph for relationship extraction.
- Python, postgresql, NVIDIA, CLI, Docker and Amazon Web Services (AWS).

12 2023 **Conference Speaker – Kisaco Research's AI-Driven Drug Discovery Summit**

- *Invited to present my research in Generative Ai at the 2023 Ai-Driven Drug Discovery Summit in Boston.*

11 2023 **Guest Lecturer – Georgetown University Genomic Data Science**

- *Invited to give lecture in generative ai applications in genomics to graduate students at Georgetown University.*

2019-Current **Data Science Projects**

2023 – Genetic Biomarker Discovery – Stage 2 Invasive Bladder Cancer
- *Statistical analysis on gene expression data from Stage 2 invasive bladder cancer*
- *Repository: https://github.com/zac-webel/Bladder-Cancer-Gene-Expression*

2023 – Tabular EHR Patient Mortality Prediction
- *I created an ensemble of shallow machine learning classifiers xgboost, random forest, and catboost to predict patient mortality given tabular EHR data.*
- *Repository: https://github.com/zac-webel/EHR-ml*

2023 – Graduate Research Paper on Federated Machine Learning
- *I co-authored a research paper on federated machine learning in healthcare applications.*
- *Paper: https://github.com/zac-webel/Federated-ML*

2023 – Imaging Artificial Intelligence
- *I created convolutional neural network that classified CT images from COVID patients.*
- Repository: https://github.com/zac-webel/CT-Scan-AI

2023 – Analyzed Differential Gene Expression for TCGA Lung Cancer Database
- R, TCGA, G-DOC HUB

2023 – DNA Copy Number Analysis for Hepatocellular Carcinoma Data
- R, TCGA

2022 – Modeling Equilibria and Stability of the Lorenz Equations
- *I created a notebook that explored the dynamics, chaos, and stability of the Lorenz Equations.*

2022 – System of Differential Equations Model – Population Dynamics - Optimization
- *Expanded SIR that modeled the effect of masks and vaccines during the COVID pandemic. Then optimized the system based on user parameters.*

2022 – Complex Visualization
- *Visualization of the contour plot, gradient and level curves of a complex multivariable function.*

2021 – Numerical Analysis Approximation – Quadrature – Interpolation – Root finding
- *I created notebooks that modeled: Orthogonal Polynomials, Least Squares, Monte Carlo Integration, Gauss-Legendre rule, Newton-Cotes quadrature, cubic splines, Hermite interpolation, Runge's function, Newton's method, and Muller's method.*

2020 – Probabilistic Modeling – A Neural Approach
- *I created a neural network that models expected yards gained on a football play resulting in a masked probability distribution that can be sampled to simulate a play.*

2019 – Tabular Data Modeling
- *Random forest model that predicted the play call using play by play data.*
- *Random forest model that predicted the blitz package for the defense on any given nfl play.*